

Science Education

JULY 2014 / VOLUME 98 / ISSUE NO 4

~~CONFIDENTIAL~~
Marygrove College Library
8425 West McNichols Road
Detroit, MI 48221

WILEY

Discover this journal online at
Wiley Online Library
wileyonlinelibrary.com/journal

Science Education

EDITOR

John L. Rudolph
Science Education
University of Wisconsin–Madison
Madison, WI 53706

Assistant to the Editor: Shannon Vakil

EDITORIAL BOARD

Glen S. Aikenhead
University of Saskatchewan
Bronwyn Bevan
Section Coeditor
Science Learning in Everyday Life
Center for Informal Learning and Schools, The Exploratorium
William J. Boone
Miami University (Ohio)
Douglas Clark
Vanderbilt University
Ravit Golan Duncan
Rutgers University
Sibel Erduran
Section Editor, *Science Studies and Science Education*
University of Limerick
Michael Ford
Section Coeditor, *Learning*
University of Pittsburgh

Jeffrey A. Greene
University of North Carolina at Chapel Hill
Leslie Herrenkohl
Section Coeditor
Science Learning in Everyday Life
University of Washington
Maria Pilar Jiménez-Alexandre
Section Coeditor
Issues and Trends
Universidade de Santiago de Compostela
Greg Kelly
Penn State University
Julie M. Kittleston
Section Coeditor
Science Teacher Education
University of Georgia
Xiufeng Liu
University of Buffalo
Jonathan Osborne
Section Coeditor
Issues and Trends
Stanford University

PAST EDITORS

Gregory J. Kelly (2006–2012)
Nancy W. Brickhouse (2001–2006)
Richard A. Duschl (1993–2001)
Leopold E. Klopfer (1979–1992)

Eileen R. Carlton Parsons
Section Coeditor
Science Education Policy
The University of North Carolina at Chapel Hill
Senay Purzer
Purdue University
Jim Ryder
University of Leeds
Troy Sadler
Section Editor, *The Books*
University of Missouri
Vic Sampson
UT-Austin
William A. Sandoval
University of California, Los Angeles
John Settlage
University of Connecticut
Marie-Claire Shanahan
University of Calgary

Sherry Southerland
Section Coeditor
Science Teacher Education
Florida State University
Jan van Driel
Leiden University
Maria Varelas
Section Coeditor, *Learning*
University of Illinois at Chicago
Carolyn Wallace
Indiana State University
Per-Olof Wickman
Stockholm University
Mark Windschitl
University of Washington
Hsin-Kai Wu
National Taiwan Normal University
Carla Zembal-Saul
Pennsylvania State University
Anat Zohar
Section Coeditor
Science Education Policy
The Hebrew University of Jerusalem

SCIENCE EDUCATION (ISSN 0036-8326 (Print); ISSN 1098-237X (Online)) is published bimonthly by Wiley Subscription Services, Inc., a Wiley Company, 111 River St., Hoboken, NJ 07030-5774. Periodical Postage Paid at Hoboken, NJ and additional offices. **Postmaster:** Send all address changes to SCIENCE EDUCATION, Journal Customer Services, John Wiley & Sons Inc., 350 Main St., Malden, MA 02148-5020.

Copyright and Copying. Copyright © 2014 Wiley Periodicals Inc. All rights reserved. No part of this publication may be reproduced, stored or transmitted in any form or by any means without the prior permission in writing from the copyright holder. Authorization to copy items for internal and personal use is granted by the copyright holder for libraries and other users registered with their local Reproduction Rights Organisation (RRO), e.g. Copyright Clearance Center (CCC), 222 Rosewood Drive, Danvers, MA 01923, USA (www.copyright.com), provided the appropriate fee is paid directly to the RRO. This consent does not extend to other kinds of copying such as copying for general distribution, for advertising or promotional purposes, for creating new collective works or for resale. Special requests should be addressed to: permissionsuk@wiley.com

Wiley's Corporate Citizenship initiative seeks to address the environmental, social, economic, and ethical challenges faced in our business and which are important to our diverse stakeholder groups. Since launching the initiative, we have focused on sharing our content with those in need, enhancing community philanthropy, reducing our carbon impact, creating global guidelines and best practices for paper use, establishing a vendor code of ethics, and engaging our colleagues and other stakeholders in our efforts. Follow our progress at www.wiley.com/go/citizenship

Information for subscribers: SCIENCE EDUCATION is published in 6 issues per year. Institutional subscription prices for 2014 are: Print & Online: US\$2107 (US), US\$2233 (Rest of World), €1456 (Europe), £1152 (UK). Prices are exclusive of tax. Asia-Pacific GST, Canadian GST/HST and European VAT will be applied at the appropriate rates. For more information on current tax rates, please go to wileyonlinelibrary.com/tax-vat. The price includes online access to the current and all online back files to January 1st 2010, where available. For other pricing options, including access information and terms and conditions, please visit wileyonline.library.com/access. **Delivery Terms and Legal Title:** Where the subscription price includes print issues and delivery is to the recipient's address, delivery terms are Delivered at Place (DAP); the recipient is responsible for paying any import duty or taxes. Title to all issues transfers FOB our shipping point, freight prepaid. We will endeavour to fulfil claims for missing or damaged copies within six months of publication, within our reasonable discretion and subject to availability. **Journal Customer Services:** For ordering information, claims and any enquiry concerning your journal subscription please go to www.wileycustomerhelp.com/ask or contact your nearest office.

Americas: Email: cs-journals@wiley.com; Tel: 1 781 388 8598 or +1 800 835 6770 (toll free in the USA & Canada). **Europe, Middle East and Africa:** Email: cs-journals@wiley.com; Tel: 44 (0) 1865 778315. **Asia Pacific:** Email: cs-journals@wiley.com; Tel: 65 6511 8000. **Japan:** For Japanese speaking support, Email: cs-japan@wiley.com; Tel: 65 6511 8010 or Tel (toll-free) 005 316 50 480. **Visit our Online Customer Get-Help** available in 6 languages at www.wileycustomerhelp.co

Back issues: Single issues from current and recent volumes are available at the current single issue price from cs-journals@wiley.com. Earlier issues may be obtained from Periodicals Service Company, 11 Main Street, Germantown, NY 12526, USA. Tel: 1 518 537 4700, Fax: 1 518 537 5899, Email: psc@periodicals.com

Advertising Sales: Inquiries concerning advertising should be forwarded to Advertising Sales Manager, Joe Tomaszewski, John Wiley & Sons, 111 River Street, Hoboken, NJ 07030: (201) 748-8895. Email: jtomaszewski@wiley.com

Production Editor: Carol Neff (cjneff@wiley.com).

Other correspondence: All other correspondence should be addressed to Science Education, Publisher, c/o John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030.

Information: For submission instructions, subscription, and all other information and to view this journal online, visit: wileyonlinelibrary.com/sce

Online Open: SCIENCE EDUCATION accepts articles for Open Access publication. Please visit <http://olabout.wiley.com/WileyCDA/Section/id-406241.html> for further information about OnlineOpen.

Disclaimer. The Publisher and Editors cannot be held responsible for errors or any consequences arising from the use of information contained in this journal; the views and opinions expressed do not necessarily reflect those of the Publisher and Editors, neither does the publication of advertisements constitute any endorsement by the Publisher and Editors of the products advertised.

This journal is indexed by CIE: *Current Index to Journals in Education* (ERIC), *Contents Pages in Education* (T&F), *Current Abstracts* (EBSCO), *Current Contents®/Social & Behavioral Sciences* (Thomson ISI), *Education Index/Abstracts* (HW Wilson), *Educational Research Abstracts Online* (T&F), *ERIC Database* (Education Resources Information Center), *FRANCIS Database* (INIST/CNRS), *IBR & IBZ: International Bibliographies of Periodical Literature* (KG Saur), *Informal Science.org* (University of Pittsburgh Center for Learning in Out-of-School Environments), *Journal Citation Reports/Social Science Edition* (Thomson ISI), *Psychological Abstracts/PsycINFO* (APA), *Research into Higher Education Abstracts* (T&F), *SCOPUS* (Elsevier), *Social Sciences Citation Index®* (Thomson ISI), *Social SciSearch®* (Thomson ISI), and *Web of Science®* (Thomson ISI). Printed in the USA by The Sheridan Group.

Science Education

CONTENTS

549/ Developing a Measure of Scientific Literacy for Middle School Students

Helenrose Fives, Wendy Huebner, Amanda S. Birnbaum, and Mark Nicolich

Published Online 18 June 2014

581/ Using Rasch Measurement for the Development and Use of Affective Assessments in Science Education Research

Toni A. Sondergeld and Carla C. Johnson

Published Online 5 June 2014

614/ Scientific Practices in Elementary Classrooms: Third-Grade Students' Scientific Explanations for Seed Structure and Function

Laura Zangori and Cory T. Forbes

Published Online 14 May 2014

ISSUES AND TRENDS

640/ Investigating the Link Between Learning Progressions and Classroom Assessment

Erin Marie Furtak, Deb Morrison, and Heidi Kroog

Published Online 18 June 2014

LEARNING

674/ Creating Opportunities for Students to Show What They Know: The Role of Scaffolding in Assessment Tasks

Hosun Kang, Jessica Thompson, and Mark Windschitl

Published Online 22 May 2014

705/ Long-Term Self-Regulation of Biology Learning Using Standard Junior High School Science Curriculum

Billie Eilam and Shoshi Reiter

Published Online 18 June 2014

THE BOOKS

738/ Trying Biology: The Scopes Trial, Textbooks, and the Antievolution Movement in American Schools

Aaron J. Sickel

Published Online 30 January 2014

740/ I Died for Beauty: Dorothy Wrinch and the Culture of Science

Gayle A. Buck

Published Online 30 January 2014

Science Education

BOARD OF REVIEWERS*

- Marianne Achiam, University of Copenhagen, Denmark
Ana Afonso, University of Minho, Portugal
Lori Andersen, Kansas State University, USA
Kevin Anderson, USA
Louise Archer, King's College London, UK
Hanna Arzi, Independent Scholar, Israel
Doris Ash, University of California, Santa Cruz, USA
Charles Ault, Jr., Lewis & Clark College, USA
Meena Balgopal, Colorado State University, USA
Tara Barnhart, University of California, Irvine, USA
Meghan Bathgate, University of Pittsburgh, USA
Orit Ben-Zvi Assaraf, Ben Gurion University of the Negev, Israel
Judith Bennett, The University of York, UK
Roland Berger, Universität Osnabrück, Germany
Adam Bertram, Monash University, Australia
Margaret Blanchard, North Carolina State University, USA
Maria Boe, Norwegian Centre for Science Education, Norway
Melissa Braaten, University of Wisconsin, USA
Muammer Çalık, Karadeniz Technical University, Fatih Faculty of Education, Turkey
Heidi Carlone, University of North Carolina-Greensboro, USA
Shu-Nu Chang Rundgren, Karlstad University, Sweden
Michael Clough, Iowa State University, USA
William Cobern, Western Michigan University, USA
Robert Danielowich, USA
Joshua Danish, Indiana University, USA
Josephine Desouza, Ball State University, USA
Niels Dohn, University of Aarhus, Denmark
Catherine Eberbach, University of Pittsburgh, USA
Maria Evagorou, University of Nicosia, Cyprus
Tessa Eysink, University of Twente, Netherlands
Noah Feinstein, University of Wisconsin-Madison, USA
Miriam Ferzli, North Carolina State University, USA
Cory Forbes, University of Iowa, USA
Danielle Ford, University of Delaware, USA
Brian Frank, MTSU, USA
Gavin Fulmer, National Institute of Education, Singapore
Mark Girod, Western Oregon University, USA
Lisa Gross, Appalachian State University, USA
Jesper Haglund, Linköping University, Sweden
Karim Hamza, Stockholm University, Sweden
Emily Harris, University of California, Davis, USA
Shusaku Horibe, University of Wisconsin-Madison, USA
Anne Hume, University of Waikato, New Zealand
Ruth Jarman, Queen's University Belfast, UK
Eugene Judson, Arizona State University, USA
Justine Kane, Wayne State University, USA
Ursula Kessels, Universitaet zu Koeln, Germany
Meredith Kier, North Carolina State University, USA
Mary Koppal, AAAS, USA
Eleni Kyza, Cyprus University of Technology, Cyprus
Douglas Larkin, Montclair State University, USA
Jari Lavonen, University of Helsinki, Finland
Norman Lederman, Illinois Institute of Technology, USA
Hee-Sun Lee, Tufts University, USA
Victor Lee, Utah State University, USA
Yew Jin Lee, National Institute of Education, Singapore
David Long, University of Kentucky, USA
Megan Luce, Stanford University, USA
Panayota Mantzicopoulos, Purdue University, USA
Eve Manz, Vanderbilt University, USA
Peter Marle, University of Colorado, USA
Christine McDonald, Griffith University, Australia
Katherine McNeill, Boston College, USA
Daniel Meyer, Illinois College, USA
Xenia Meyer, University of California at Berkeley, USA
Alandeom Oliveira, State University of New York, University at Albany, USA
Leif Östman, Department of Education, Sweden
Nicos Papadouris, University of Cyprus, Cyprus
Krystal Perkins, University of West Georgia, USA
Erin Peters Burton, George Mason University, USA
Lilian Pozzer-Ardenghi, University of Manitoba, Canada
Leonie Rennie, Curtin University, Australia
Vincent Richard, Université Laval, Canada
Kelly Riedinger, University of North Carolina, Wilmington, USA
Catherine Rieggle-Crumb, University of Texas, USA
Ann Rivet, Teachers College Columbia University, USA
Gillian Roehrig, University of Minnesota, USA
David Rudge, Western Michigan University, USA
Maria Araceli Ruiz-Primo, University of Colorado Denver, USA
Rosemary Russ, University of Wisconsin-Madison, USA
Suna Ryu, University of California at Los Angeles, USA
Ala Samarapungavan, Purdue University, USA

*Individuals who complete two or more reviews during a calendar year are included on the Board of Reviewers.

BOARD OF REVIEWERS

Silvana Santos, Universidade Estadual da Paraíba,
Brazil

Renée Schwartz, Western Michigan University,
USA

Steven Semken, Arizona State University, USA

Ji Shen, University of Miami, USA

Harvey Siegel, University of Miami, USA

Tiffany-Rose Sikorski, The George Washington
University, USA

Carol Smith, University of Massachusetts, USA

Leigh Smith, Brigham Young University, USA

Shawn Stevens, University of Michigan, USA

Keith Taber, University of Cambridge, UK

Robert Tai, University of Virginia, USA

Tali Tal, Technion, Israel

Edna Tan, MSU, USA

Hsiao-Lin Tuan, National Changhua University of
Education, Taiwan

Esther van Dijk, University of Hildesheim, Germany

Grady Venville, University of Western Australia,
Australia

Mihye Won, Curtin University, Australia

David Wong, Michigan State University, USA

Anat Yarden, Weizmann Institute of Science, Israel

Larry Yore, University of Victoria, Canada

Zacharias Zacharia, University of Cyprus, Cyprus

Gabor Zemlen, Budapest University of Technology
and Economics, Hungary

Heather Zimmerman, Pennsylvania State University,
USA

Science
Education

Developing a Measure of Scientific Literacy for Middle School Students

HELENROSE FIVES,¹ WENDY HUEBNER,² AMANDA S. BIRNBAUM,²
MARK NICOLICH³

¹Department of Educational Foundations and ³Department of Health and Nutrition Sciences, Montclair State University, Montclair, NJ 07043, USA; ²Cogimet, Lambertville, NJ 08530, USA

Received 15 July 2013; accepted 10 March 2014

DOI 10.1002/sce.21115

Published online 18 June 2014 in Wiley Online Library (wileyonlinelibrary.com).

ABSTRACT: Scientific literacy reflects “a broad and functional understanding of science for general education purposes” (DeBoer, 2000, p. 594). Herein, we present the ongoing development of the Scientific Literacy Assessment (SLA), a work-in-progress measure to assess middle school students’ (ages 11–14) scientific literacy. The SLA includes a selected response measure of students’ *demonstrated* scientific literacy (SLA-D) and a motivation and beliefs scale based on existing measures of self-efficacy, subjective task value, and personal epistemology for science (SLA-MB). Our theoretical conceptualization of scientific literacy guided the development of our measure. We provide details from three studies: Pilot Study 1 ($n = 124$) and Pilot Study 2 ($n = 220$) describe the development of the SLA-D by conducting iterative item analyses of the student responses, think-aloud interviews with six students, and external expert feedback on the items in the SLA-D. Study 3 describes the testing of our prototype measure ($n = 264$). We present a validity argument including reliability evidence that supports the use of the current version of the SLA to provide evaluation of middle school students’ scientific literacy. Our resulting SLA includes the SLA-D in two versions, each with 26 items and the SLA-MB with 25 items across three scales: value of science, scientific literacy self-efficacy, and personal epistemology. © 2014 Wiley Periodicals, Inc. *Sci Ed* 98:549–580, 2014

Correspondence to: Helenrose Fives; e-mail: fivesh@mail.montclair.edu

Contract grant sponsor: Science Education Partnership Award (SEPA), supported by the National Center for Research Resources.

Contract grant sponsor: Division of Program Coordination, Planning, and Strategic Initiatives of the National Institutes of Health.

Contract grant number: 8R25 OD011117-05.

Supporting Information is available in the online issue at wileyonlinelibrary.com.

© 2014 Wiley Periodicals, Inc.

DEVELOPING A MEASURE OF SCIENTIFIC LITERACY FOR MIDDLE SCHOOL STUDENTS

At its core, scientific inquiry is the same in all fields. Scientific research, whether in education, physics, anthropology, molecular biology, or economics, is a continual process of rigorous reasoning supported by a dynamic interplay among methods, theories, and findings. It builds understanding in the form of models or theories that can be tested. (Scientific Research in Education; National Research Council [NRC], 2002, p. 2)

Scientific literacy is the ability to understand scientific processes and to engage meaningfully with scientific information available in daily life. Meaningful learning is understood as the connection of new information with prior knowledge in personally relevant ways (e.g., Aikenhead, 2011; Ausubel, 1977; Berry, Loughran, & Mulhall, 2007). Thus, we see scientific literacy as “a broad and functional understanding of science for general education purposes and not preparation for specific scientific and technical careers”; this functionality refers to the ability to use science to “live more effectively with respect to the natural world” (DeBoer, 2000, p. 594). This definition draws on perspectives from multiple sources including research and policy documents (NRC, 1996, 2012; Organisation for Economic Co-operation and Development [OECD], 2007) and science education researchers (e.g., Bybee, 2008; DeBoer, 2000; Laugksch, 2000; Roberts, 2007). There is no single accepted definition of scientific literacy; rather, the many characterizations of scientific literacy discussed in the literature include varying elements of competencies in science inquiry, content knowledge, and attitudes toward science (e.g., DeBoer, 2000; Roberts, 2007). Trends in science education policy have emphasized the importance of scientific literacy as a transferable outcome of science education. Several measures of scientific literacy currently exist (e.g., Bybee, 2008; OECD, 2006; Wenning, 2006, 2007). However, none of these target middle school students (aged 11–14 years, Grades 6–8 in the United States). Furthermore, most current measures draw on some degree of complex knowledge of one or more specific science field/disciplines and most measures do not include assessment of attitudes toward science. The purpose of our investigation was to develop a measure to assess the scientific literacy of middle school students that included assessments of the ability to think scientifically and students’ motivation and beliefs toward science while being as field/discipline general as possible.

In designing this measure of scientific literacy, we sought to achieve a degree of science field (e.g., life, physical)/discipline (e.g., biology, astronomy) generality. That is, we attempted to measure the aspects of scientific literacy identified in our framework (described below) in ways that did not rely on field/discipline scientific knowledge (e.g., photosynthesis, simple machines, atomic structure); rather, we focused on the processes of science that span specific fields/disciplines. In part, this decision is based on the recognition that middle school students may not have a common depth of field/discipline knowledge from which to draw; thus, our goal was an instrument that can be used broadly to make valid inferences and evaluations by educators and educational researchers.

SCIENTIFIC LITERACY: THE CONCERN FOR SCIENCE EDUCATION

Educators and policy makers have made repeated calls for improved K-12 science education and defined performance expectations to reinforce the need for science as inquiry to improve scientific literacy (American Association for Advancement of Science [AAAS], 1993; National Assessment Governing Board [NAGB], 2010; NRC, 1996, 2012; OECD, 2007). These expectations are based on the premise that science is a recursive, dynamic process of asking questions, investigating, and then asking more questions, and that

these approaches can better engage children, who are naturally curious and learn through experience.

Most recently, in the United States, the National Research Council's report entitled *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* stated that hands-on investigative science is crucial in that it "gives [students] an appreciation of the wide range of approaches that are used to investigate, model, and explain the world" (NRC, 2012, p. 42). Moreover, authentic experiences in science can begin at a young age; the NRC (1996) report claimed that "[s]cience literacy begins with attitudes and values established in the earliest years . . ." (p. 18) and ". . . attitudes and values established toward science in the early years will shape a person's development of scientific literacy as an adult" (p. 22). This requires that education provide learners with a science curriculum that will facilitate the development of scientific literacy. We posit that given the goal of improved scientific literacy among students, we need to fill a gap in our ability to measure such scientific literacy.

Our first step was to develop an up-to-date conceptualization of the nature of scientific literacy reflective of extant theory in the field (described in the next section). This conceptualization framed the development of our measure. Findings from our two rounds of iterative testing of the measure informed our final conceptualization of scientific literacy. The pilot testing and final prototype testing are described later in the manuscript followed by a presentation of our findings and recommendations for use.

THEORETICAL REVIEW: CONCEPTUALIZATION OF SCIENTIFIC LITERACY

We engaged in a systematic review of the literature on scientific literacy that focused on how scientific literacy is defined, the various components that have been identified, and previous measures used to assess it. Reviews by Laugksch (2000), DeBoer (2000), Dillon (2009), Holbrook and Rannikmae (2009), and Roberts (2007) provided essential historical context for understanding the development of the concept of scientific literacy over the last 50 years. We also reviewed policy documents from leading science education agencies (e.g., AAAS, 1993; National Science Teachers Association [NSTA], 1991; NRC, 1996, 2012; OECD, 2007) to identify core capabilities that are considered essential to scientific literacy. Each definition or capability list was broken down into specific capabilities and compared across documents to identify components for assessment. We initially generated 12 components of scientific literacy to assess. We independently engaged in a theoretical analysis of these components to synthesize and better reflect the field. We then compared and discussed the individual syntheses and grouped similar components together, resulting in a total of six components comprising our initial framework.

While our perspective on scientific literacy is informed by the extant literature (described below), we focus more on the processes of science that span specific fields of study and disciplines. While some scholars contend that an information-rich knowledge of science is necessary for true scientific literacy (e.g., Shamos, 1995), others emphasize scientific literacy as active participation in the sociocultural potential and consequences of science (e.g., Cross & Price, 1992) that could lead to social activism (Hodson, 1999), and still others define scientific literacy as the "ability to deal with science in the news" (Hazen & Trefil, 1991, p. xii). The perspective we take attempts to find a middle ground that recognizes scientific literacy as knowledge of the nature of the field and its processes so that one can engage (in whatever form that takes for the individual) with science pragmatically and meaningfully in daily life. Our framework for conceptualizing scientific literacy is presented in Table 1 with a summary of supporting references. This initial framework

TABLE 1
Summary of Initial Scientific Literacy Construct and Components

Components	Supporting Literature												
	Showalter (1974)	Shen (1975)	Arons (1983)	Miller (1983)	AAAS (1993)	Hazen and Trefil (1991)	NSTA (1991)	NRC (1996)	DeBoer (2000)	Duit and Treagust (2003)	OECD (2007)	Holbrook and Rannikmae (2009)	NAGB (2010)
Role of science: Identify questions that can be answered through scientific investigation; understand the nature of scientific endeavors; understand generic science concepts	✓	✓	✓	✓				✓		✓	✓		
Scientific thinking and doing: Describe natural phenomena; recognize patterns; identify study variables; ask critical questions about study design; reach/evaluate conclusions based on evidence	✓		✓		✓		✓	✓	✓	✓	✓		✓
Science and society: Apply scientific conclusions to daily life; identify scientific issues underlying policy decisions; understand the role of science in decision making	✓	✓		✓	✓	✓	✓	✓	✓		✓		
Science media literacy: Develop questions to assess the validity of scientific reports; question the sources of science reporting ^a								✓	✓				
Mathematics in science: Use mathematics in science; understand the application of mathematics in science					✓								
Science motivation and beliefs: Value of science; self-efficacy for scientific literacy; personal epistemology of science	✓	✓	✓				✓		✓	✓	✓	✓	

^aAfter our pilot studies, we subsumed science media literacy into the science and society component.

included six components that together reflect our perspective on the nature of scientific literacy: role of science, scientific thinking and doing, science and society, science media literacy, mathematics in science, and science motivation and beliefs. In the sections that follow, we briefly describe each of these components and how each is reflective of scientific literacy.

Role of Science

The first component in our framework, role of science, reflects the way that science can function in terms of understanding (a) the kinds of questions that can be answered through science, (b) the nature of scientific activities, and (c) generic scientific concepts present across field/discipline areas (e.g., variables, experiment, observation, etc.). As indicated in Table 1, seven of the 13 resource frameworks of scientific literacy included the ability to identify scientific questions as part of their conceptualization (i.e., Arons, 1983; Duit & Treagust, 2003; Miller, 1983; NRC, 1996; OECD, 2007; Shen, 1975; Showalter, 1974). The ability to recognize scientifically investigable questions (Duit & Treagust, 2003) provides some access to individuals' understanding of the nature of science, scientific methods, and what counts as evidence in science. A scientifically literate person, at the very least, must be able to determine whether and how science can be used to address questions in daily life. For instance, when shopping for a car, the scientifically literate person can determine what details provided by the salesperson are scientifically verifiable (e.g., fuel consumption, safety ratings) and which are issues of personal preference (e.g., interior color, prestige).

Scientific Thinking and Doing

While the first component emphasized the ability to recognize when science can be used to answer questions, this second component refers to actually *doing* the science needed to answer those questions. Thus, the scientifically literate are able to engage observational and analytical processes that are required for scientific thought. Scientific thinking includes the abilities to “describe, explain, and predict natural phenomenon” (NRC, 1996, p. 22), generate and evaluate scientific evidence (NRC, 1996), understand the difference between inference and observation (Arons, 1983), and identify patterns in data (NAGB, 2010). Thus, this component refers to the ability to design and conduct studies to address questions that can be answered by science.

In addition, this component includes the ability to question scientific methods, use evidence to support or refute arguments, and apply evidence-based conclusions (Duit & Treagust, 2003; NAGB, 2010; NRC, 1996; NSTA, 1991). The individual's ability to understand and apply scientific methods well enough to question and critique those methods when presented, to evaluate types of evidence offered in light of a research design, and to make conclusions about the findings presented are also included in this component. Furthermore, the abilities to apply what one knows about science (methods, theories, etc.) to new scientific endeavors and to evaluate those new endeavors through scientific reflection are also conceived to be part of scientific thinking.

Science and Society

The abilities to identify (a) scientific issues underlying local, national, and international policy and (b) science in decision-making processes are reflected in this component. We argue for a broad perspective on policy to include decisions made in the individual's home, workplace, or school, to larger contexts of town, state, national, and international arenas.

This component refers to the individual's ability to both recognize the role of science in decision making as well as the reasons for why science may not always be the deciding factor on policy decisions. The interaction between science and society was also noted by several other frameworks for scientific literacy (i.e., AAAS, 1993; DeBoer, 2000; Hazen & Trefil, 1991; Liu, 2009; Miller, 1983; NRC, 1996; NSTA, 1991; OECD, 2007; Shen, 1975; Shonwalter, 1974). Similar to our conceptualization, these other researchers described the need for people to develop a balanced perspective that integrates scientific thinking with social norms and ethical values (e.g., Miller, 1983; NSTA, 1991).

Science Media Literacy

Scientific media literacy refers to the individual's ability to critique scientific findings described or portrayed in the popular media and is closely related to the previous component in practice (Jarman & McClune, 2007). This includes the ability to develop questions to assess the validity of scientific reporting found in news reports or other media outlets and to question the sources of evidence provided for alternative goals or priorities. There is some precedence for science media literacy from other frameworks of scientific literacy that recognized this component as a unique aspect of scientific literacy (i.e., DeBoer, 2000; NRC, 1996). Furthermore, akin to the NRC (1996) *National Science Education Standards*, and the *Beyond 2000: Science Education for the Future* report from the United Kingdom (Millar & Osborne, 1998), we see a distinction between the ability to read and understand scientific reports and the ability to recognize the need to engage that scientific thinking when exposed to popular media. Similarly, DeBoer (2000) argued for the development of citizens who are able to "critically follow reports and discussion about science that appear in the media . . ." and recognize the direct role that science has in daily life (p. 592). This component recognizes the need for the individual to be able to activate scientific thinking when necessary and apply that thinking to information presented in the "normal" world of the person. Finally, as individuals move into adult life, the source of ongoing scientific literacy is often the news media and with this source comes embedded biases and persuasive tactics. Science media literacy seeks to facilitate the ongoing learning of science throughout adulthood with the ability to be a critical consumer of that information (Jarman & McClune, 2007; Zimmerman, Bisanz, Bisanz, Klein, & Klein, 2001). The current communications environment exposes individuals to constant information and arguments; being an informed citizen requires the ability to critically, quickly, and accurately ascertain the basis for arguments from among obviously scientific arguments about climate change to the more subversive use of science seen in advertisements for weight loss supplements and devices. Note that in our final conceptualization of scientific literacy we collapse science and society with science media literacy into one component.

Mathematics in Science

Our framework includes mathematics in science as a distinct component within scientific literacy. The inclusion of this component is supported by the AAAS (1989, 1993) and others in the field (e.g., Hamm, 1992; Yore, Pim, & Tuan, 2007). Domain-specific literacy may be described as *fundamental*, able to engage in domain-specific discourse, and *derived*, having an understanding of the content in the domain (Norris & Phillips, 2003). Yore and colleagues (2007) used this distinction to draw a parallel between literacy in the domains of mathematics and science and argued that literacy within either of these fields would require an interaction of both fundamental and derived literacies. The importance of literacy in both mathematics and science is underscored by the assessment of these separate literacies

by the Programme for International Student Assessment (PISA), which considers literacy to include the ability to apply knowledge across disciplines (OECD, 2003).

While we agree with the importance of mathematical literacy for reasons similar to the need for scientific literacy, in our framework, akin to that offered by AAAS (1989), we attempt to identify the kinds of mathematical understandings that are inherent to evaluating scientific findings. A working knowledge of mathematics as used in science (e.g., graph reading and the understanding of proportions and percentages) is necessary to fully understand science in everyday life; we consider this different from basic computation. The use of statistics and visual representations of numerical data has become commonplace in U.S. media. Everything from the representation of the number of search results as “O’s” in Google to the forecasting of tomorrow’s weather is reported through mathematics and visual representations of that mathematics. As such, an understanding of the mathematics that is used to communicate scientific findings and results is required to recognize or critically examine and understand science in media or understand issues of science in society.

Science Motivation and Beliefs

More than knowledge is needed to be a scientifically literate person; one must also have the motivation and beliefs necessary to engage that knowledge when needed as part of one’s daily life. Therefore, we chose to include components related to students’ motivation for and beliefs about science. Attitudes, values, and beliefs have been identified by others as components of scientific literacy (e.g., AAAS, 1989; Arons, 1983; Holbrook & Rannikmae, 2007; NRC, 1996; NSTA, 1991; OECD, 2007; Ryder, 2001; Shen, 1975). Despite the recommendation to address values and beliefs in conceptions of scientific literacy, little effort has been made to articulate just what those values and beliefs should be. To engage in an analysis of that nature was beyond the scope of our current work. However, we felt that to ignore this aspect of scientific literacy entirely would be disingenuous to a modern understanding of this concept. We were guided by Gauld’s (1982) description of the “scientific attitude” as the motivation needed to convert knowledge and skills into scientific procedures and engagement. From this perspective, we reviewed some of the work on students’ motivation and beliefs in science and selected three constructs relevant to the successful engagement of scientific literacy: value (subjective task value), confidence (self-efficacy), and beliefs about knowledge and knowing (personal epistemology). Therefore, we perceive a scientifically literate person as one who values science (intrinsically and for utility purposes: Wigfield & Eccles, 2000), feels capable of engaging in scientific activities (self-efficacy: Ketelhut, 2010), and believes that knowledge in science is developed by humans and is changing (personal epistemology: Conely, Pintrich, Vekiri, & Harrison, 2004).

Motivation researchers have examined the relations between the value students attribute to content area or achievement tasks and their engagement and achievement in school (e.g., Bøe, 2012; Eccles & Wigfield, 2002). Subjective task value is used to describe the value that learners have for academic tasks and has been described in four ways: *intrinsic value* refers to learners’ experiences of “fun” or enjoyment for the task itself, *attainment value* refers to how important success on a task is for the learner’s sense of self, *utility value* occurs when the learner sees the task as useful for some other goal, and the last area is *cost* that refers to what a learner must give up to engage in the task (Eccles, Barber, & Jozefowicz, 1999; Wigfield & Eccles, 2000). Task value is salient to scientific literacy. If a learner is to engage in his/her scientific literacy as part of daily life, then they must see some value for doing so, either because they enjoy it, they see themselves as the kind of person to think scientifically, or they see it as useful.

In the field of achievement motivation, self-efficacy beliefs are identified as beliefs held by an individual about his/her ability to organize and execute acts to bring about the desired outcome (Bandura, 1997). In other words, this refers to their perceived confidence in completing tasks. One definition of scientific literacy explicitly addressed the importance of feelings of self-efficacy in terms of confidence. Scientific literacy was described as

[t]he capability to function with understanding and *confidence*, and at appropriate levels, in ways that bring about empowerment in the made world and in the world of scientific and technological ideas. (emphasis added, UNESCO, 1993, p. 15)

This definition and others indicated that it is not enough for students to be able to know about science or how to engage in science but that they must actually do so and feel confident about that capability, that is, they must have self-efficacy for science. Self-efficacy beliefs influence learners' choices, effort, and persistence, and routinely predict academic achievement (e.g., Britner & Pajares, 2001; Bryan, Glynn, & Kittleson, 2011; Lent, Brown, & Gore, 1997; Pajares & Valiante, 1997; Shell, Colvin, & Bruning, 1995). Thus, we chose to include the construct of self-efficacy for engaging in activities associated with scientific literacy.

Personal epistemology refers to individuals' domain-specific beliefs about knowledge and knowing (Hofer & Pintrich, 1997). Hofer's (2000) epistemological theories perspective suggests that beliefs about knowledge and knowing serve as interconnected theories that learners use as they engage with content and the world. Specifically, there are beliefs about "the nature of knowledge (what one believes knowledge is)" and beliefs about "the nature or process of knowing (how one comes to know)" (Hofer, 2000, p. 361). Within each of these frames, two dimensions of beliefs have been identified. Beliefs about the nature of knowledge have been described along two continua: certainty (knowledge is certain vs. knowledge is fluid) and simplicity (knowledge is made up of discrete separate units vs. knowledge is integrated and complex). Beliefs about knowing are described as beliefs about the source of knowledge (from authority or outside the person vs. constructed by individuals) and the justification of knowledge. Students' epistemological beliefs influence learning outcomes (e.g., Perkins, Jay, & Tishman, 1993; Songer and Linn, 1991), strategic processing, and reading comprehension (e.g., Braten, Stromoso, & Samuelson, 2008).

Limitations in Our Conceptualization of Scientific Literacy

We constrained our conceptualization of scientific literacy to what we felt could be tested in middle school students through paper and pencil measures. We appreciated calls from DeBoer (2000) and Holbrook and Rannikmae (2009) to maintain an open-ended and situation/culturally specific conceptualization of scientific literacy and we agree that the construct of scientific literacy includes fluid situation-specific applications of science in daily life. For example, Duit and Treagust (2003) argued for the inclusion of *collaboration* in a conception of scientific literacy, referring to individuals' abilities to interact with each other as they engage in scientific inquiry. However, we omitted such conceptions of scientific literacy from our framework, not because they are not valued components of this construct, but because we felt that these conceptions required more nuanced performance-based assessments to accurately assess individuals' abilities in these areas. Thus, we recognize that our conceptualization of scientific literacy is limited to those components we felt could be appropriately assessed through the type of measure we wanted to design.

MEASURING SCIENTIFIC LITERACY

Considering the importance of science literacy outcomes, it is surprising to discover the paucity of available measures that attempt to assess it. For example, in the United States, many state and national standardized tests for students attempt to measure scientific literacy, yet the scope of such tests (and the classroom instruction that precede them) is so broad that teachers engage in surface level coverage of a wide range of topics at the cost of allowing students time to focus deeply and learn a few central scientific concepts (Lambert, 2006). In a similar vein, it is recognized that most U.S. state and national testing programs do not address abilities in scientific inquiry (Fuchs, 2008). Several measures of scientific literacy currently exist (Bybee, 2008; Laugksch & Spargo, 1996; Liu, 2009; OECD, 2007; Wenning, 2007). However, existing measures have three key limitations in that they (1) tend to be field/discipline specific, (2) are intended for students at the secondary or university levels, and (3) ignore the assessment of students' motivation for and beliefs about science.

Field/Discipline Specificity

Items on existing measures of scientific literacy are largely information-dependent. By that we mean that learners must have sufficient scientific information (Jenkins, 2003) to respond accurately to test items. In contrast, scientific literacy should emphasize those aspects of science that transcend specific fields/disciplines, focus on the processes of science, and reflect scientific training (Jenkins, 2003). Science field/discipline-specific measures of scientific literacy are evidenced in the PISA science literacy measure that utilizes the *environment and natural resources* as appropriate context for measuring scientific literacy among 15-year-olds in 57 countries (Bybee, 2008). Similarly, the test by Wenning (2007) emphasized understanding of *physics*, with some of the items focused on general scientific thinking.

We agree that there is an important place for assessment of student understanding of specific science topics or information; however, we were interested in a measure of "their understanding of science as an approach" (NRC, 2012, p. 263). The NRC (2012) proposed a framework for K-12 science education and standards that defined eight science *practices* (e.g., asking scientific questions, engaging in argument from evidence) and seven *cross-cutting concepts* (e.g., pattern recognition, identifying cause and effect relations) that span fields/disciplines and are reflective of scientific literacy from our perspective (NRC, 2012). We agree with the premise of this recent framework for science education that "[a]lthough the practices used to develop scientific theories . . . differ from one science domain to another, all sciences share certain common features at the core of their inquiry-based and problem-solving approaches" (NRC, 2012, p. 26).

In this way, we, perhaps, sidestep one common tension in the discourse around scientific literacy, the tension between content-focused and issues-focused science (DeBoer, 2000; Roberts, 2007). Roberts (2007) framed this tension as revealed in two "visions" of scientific literacy (p. 730). Vision I reflected a focus on the knowledge of science from within the discipline, and emphasized the importance of knowledge of scientific findings, principles, and laws as a basis for engagement with the field. In contrast, Vision II garnered its focus from the issues and experiences of daily life that hold within them a scientific component. Taking each of these perspectives to extreme outcomes suggests that in Vision I only expert scientists can ever become *truly* scientifically literate; it is only with vast knowledge of the content and process of the domain that one can converse and understand fully the meaning of scientific discourses (Shamos, 1995). Similarly, a Vision II extreme perspective may lead to a conception of scientific literacy as merely functional, an ability to engage

superficially with science and to understand when it is applicable to daily life (Shamos, 1995). To use a metaphor, Vision I literacy would be the equivalent of knowing a foreign language well enough to write, produce, create, appreciate, and consume literature in that language, whereas Vision II literacy would be equivalent to conversational language that would facilitate navigating the local areas, communicating to purchase goods, and finding directions.

In his seminal review of the literature on scientific literacy, Roberts (2007) stated “[t]here is no consensus about the meaning, or even the constituent parts, of SL [scientific literacy] – with one exception: everyone agrees that students can’t be scientifically literate if they don’t know any science subject matter” (p. 735). This conclusion highlights the inherent challenges in assessing scientific literacy, to which our work responds in two key ways. First, given the lack of consensus, we offer a definition and framework for scientific literacy that is operationalized by the measure we constructed and tested. We understand that this definition is limited and may be contested on theoretical grounds and sociopolitical goals. Thus, we offer *one* way to assess scientific literacy. Second, the content assessed in this test is intended to be relevant across the fields and disciplines of science. Our test is designed to assess middle school students’ ability to recognize the underlying science processes and concerns at issue across a range of fields/disciplines. Success on this test should rest on the learners’ understanding of scientific processes rather than recall of information from different disciplines of science.

We believe it is imperative to develop measures to assess whether or not principles of science, critical thinking, and problem solving are being effectively taught and learned. In our development of field/discipline-general items, we recognized that any particular item would have science content topics in it, and if the person responding to the item has prior knowledge of that topic, he/she will most likely perform better on that item (e.g., Alexander, Kulikowich, & Schultze, 1994). Thus, in our efforts to address this, we sought to vary the science topics across the items to emphasize everyday examples.

Middle School Level

Despite the work that has been done in the field to develop tools to measure scientific literacy, this work has not addressed the specific needs of middle school students. This age group requires a tool to assess their specific abilities and needs for three reasons. First, the middle school period marks the preparation for secondary education in the United States and as such there are frequently changes in how science is taught and by whom. It is usually marked by a change from science being taught by a classroom teacher who offers multiple subjects to a dedicated science teacher with content area expertise. A tool for assessing scientific literacy at this juncture of a student’s academic career can provide meaningful information for classroom teachers as a possible formative assessment and researchers targeting science education to promote scientific literacy. Second, and in conjunction with the previous reason, science education starting at Grade 7 is a typical and entrenched academic subject worldwide (Holbrook & Rannikemae, 2007). Thus, we can, with some certainty, argue that formal instruction in science taught by science experts is offered starting in sixth grade (around age 11). The amount of variability in instruction, content, and expertise prior to Grade 7 is potentially very high. Therefore, targeting this tool for Grades 6–8 (11–14 years old) provides a good baseline for possible future development. Finally, students tend to report less interest or value for school subjects in general (e.g., Wigfield & Eccles, 2000) and science in particular (e.g., Osborne, Simon, & Collins, 2003) as they transition from elementary to secondary school. Targeting this measure for middle school allows for the assessment of the relation among knowledge, motivation, and beliefs

at this known transitional time, as well as possible predictors or facilitators of scientific literacy.

Motivation and Beliefs in Scientific Literacy Assessment

Existing measures of scientific literacy do not assess motivation and beliefs in science despite the theoretical call from scholars and organizations for these perspectives to be included in the conception of a person who is scientifically literate (Arons, 1983; DeBoer, 2000; Holbrook & Rannikmae, 2007; NSTA, 1991; Shen, 1975; Showalter, 1974). We identified task value, self-efficacy, and personal epistemology as salient motivation and belief constructs for including in a measure of scientific literacy. Together these constructs tap into the value individuals hold for science, their confidence to engage in science, and their belief in the nature of science knowledge.

AIMS OF THE INVESTIGATION

Our overarching aim was to develop and test the Scientific Literacy Assessment (SLA) measure that would allow researchers and educators to make valid inferences about middle school students' scientific literacy. To achieve this goal, we developed two sets of measures to be administered together in a single instrument. The SLA-D assesses *demonstrated* scientific literacy through a series of multiple-choice items that use everyday situations and examples, rather than field/discipline-specific scientific knowledge, to test scientific literacy through the examination of understandings of the role of science, scientific thinking and doing, science and society, science media literacy, and mathematics in science. The other component of the SLA, the SLA-MB, assesses *motivation and beliefs* associated with scientific literacy. The SLA-MB includes three adaptations of three previously developed Likert-type scales to assess students' motivations and beliefs in relation to science.

OVERVIEW OF METHODOLOGY: A MULTISTAGE APPROACH TO MEASURE DEVELOPMENT

Our approach to the design and development of this measure was informed by the unitary construct of validity advocated by the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], National Council on Measurement in Education [NCME], 1999). The unitary construct of validity recognizes a variety of types of evidence that may support a validity argument but that "validity involves an overall evaluation of the plausibility of the intended interpretations," that is validity is a property of the inferences made and not the measure itself (Kane, 1994, p. 136). The *Standards* put forth five types of evidence to use in supporting a validity argument for the use of a measurement tool; these include evidence based on test content, response process, internal structure, relations to other variables, and consequences of testing (AERA, APA, NCME, 1999). As Kane (2012) recently wrote "[t]he kinds of evidence required for validation are determined by the claims being made . . ." (p. 3). We argue that the SLA will provide users with data that can be used to make valid inferences about individuals' demonstrated scientific literacy (SLA-D) and scientific literacy motivation and beliefs (SLA-MB). With these goals in mind, we emphasized in our measure development evidence based on *test content*, *response process*, and *internal structure*. Table 2 overviews our three-study validation process and the types of validity evidence we offer to support the use of this tool.

TABLE 2
Validation Process and Sources of Validity Evidence

Overview of Validation Process and Sources of Validity					
Measure Details			Sources of Validity Evidence		
SLA Measure	Item Type	Intended Inferences	Test Content	Response Process	Internal Structure
SLA-D Study 1	57 multiple-choice (MC) items	Higher scores on this measure indicate stronger demonstrated scientific literacy	<ul style="list-style-type: none">▪ Alignment of items to definition scientific literacy• Readability statistics▪ Well written items, follow item writing guidelines• Construction and review of items by interdisciplinary test construction team• Review of items by SEPA colleagues and four science teachers in Study 1	Think aloud four students	<ul style="list-style-type: none">• Item analysis• Discrimination index• Item-total correlations
SLA-D Study 2	53 MC items			Think aloud two students	
SLA-D Prototype Study 3 (two versions)	26 MC items			No think aloud	<i>Same as above and</i> <ul style="list-style-type: none">• Kuder—Richardson 20• Factor analysis
SLA-MB Prototype Study 3	25 Likert scale items	Higher scores on these subscales indicate motivation and beliefs reflective of a scientifically literate person	Alignment of items to definition of scientific literacy motivation and beliefs		<ul style="list-style-type: none">▪ Cronbach's alpha▪ Factor analysis

Evidence Based on Test Content

To construct the SLA-D, we engaged in an iterative process of selected response item (i.e., multiple choice) item generation, evidence gathering, and redesign that fostered the development of the measure. Throughout our process, we relied on *evidence based on test content* as recommended in the *Standards* (AERA, APA, NCME, 1999), using the following strategies: (1) developing items specifically to align with each of the components identified in our framework; (2) discussion among our interdisciplinary research team composed of two epidemiologists (one who has extensive experience implementing inquiry science in K-12 education), a statistician, and an educational psychologist (who was a middle school science teacher for 6 years); (3) subjecting our initial measure to evaluation by experts in science education; and (4) crafting the items following the item writing recommendations of Haladyna, Downing, and Rodriguez (2002).

To develop a pool of candidate items, we reviewed existing measures to identify items to include or adapt, and also developed original (*de novo*) items. For the *de novo* items, each member of the research team initially drafted multiple items aligned to specific components of our framework of demonstrated scientific literacy. We wrote all items at or below the sixth-grade level according to the Flesch–Kincaid index provided by Microsoft Word. We did this because we did not want reading ability to confound our assessment of scientific literacy for this measure. We do recognize, however, that differences in reading ability and language fluency will impact individual student's scores on this, or any, measure of scientific literacy. Furthermore, the Flesch–Kincaid index does not take into account what the reader brings to the task of reading such as prior knowledge of the subject matter or interest in the topic and factors that can influence reading comprehension (e.g., Alexander et al., 1994; Baldwin, Peleg-Burckner, & McClintock, 1985).

All items were passed on in a “merry-go-round” fashion so each team member reviewed all items constructed, making changes as needed, and including new or revised items. As described below, we ran several empirical trials of the SLA-D; after each trial, we reviewed all items, distractors, and responses as a team, and engaged in collaborative discussion and review to revise, delete, or generate new items as needed. At the end of our testing processes, most of the final SLA-D items had been developed by our team (see Appendices 1 and 2 in the Supporting Information), with a total of six items derived (and used with permission) from AAAS Project 2061 (2011) and Dillashaw and Okey (1980).

For the assessment of scientific literacy motivation and beliefs (SLA-MB), we followed the same strategy of aligning our description of scientific literacy to items for inclusion in our measure described above. We examined the existing measures in the field and compared them to the definition of this component in our theoretical framework and identified three existing measures for use and adaptation. The measures were selected based on the theoretical basis of their development and congruence with the motivation and beliefs identified as salient for scientific literacy. We selected Wigfield and Eccles' (2000) measure of achievement value to assess each of the three achievement values for science: intrinsic value (fun), attainment value (importance), and utility value (usefulness). Kettlehut's (2010) measure of self-efficacy for scientific inquiry was adapted to measure students' perceptions of capability for engaging in activities reflective of scientific literacy. Ketelhut developed this tool from a sound theoretical base and engaged in measure development that provided ample validity evidence to support the use of this scale for the intended purpose. To assess students' beliefs about the source and certainty of knowledge in science that seemed most connected to issues of scientific literacy, such that they underscore the nature of science as an evolving domain with multiple responses to questions in the field, we used the two subscales from Conley and colleagues (2004).

Evidence Based on Response Process

In addition to validity evidence based on test content, we also gathered evidence for the SLA-D based on response processes by performing think-aloud interviews (e.g., Presser et al., 2004) with six middle school students (Pilot Studies 1 and 2).

Evidence Based on Internal Structure

We collected data from middle school students (Pilot Studies 1 and 2; Prototype Study 3) for evidence based on internal structure by examining responses to the SLA-D using statistical analysis of each item and of the overall measure. The scales selected for the SLA-MB had previously demonstrated evidence of internal structure and reliability. Wigfield and Eccles' (2000) task value measure has demonstrated sound reliability. Kettlehut (2010) tested her self-efficacy scale with 2,000 middle school students and reported a Cronbach's alpha of .86. The personal epistemology scales were used with fifth-grade students to assess their beliefs about science at two time periods and demonstrated acceptable reliability (i.e., the source of scientific knowledge: $\alpha = .81, .82$; the certainty of knowledge in science: $\alpha = .78, .79$; Conley et al., 2004). In Prototype Study 3, we evaluated the factor structure of the motivation and beliefs measure using principal components factor analysis and examined the reliability evidence for these scales. A brief overview of the three studies, research questions, procedures, and findings can be found in Table 3.

PILOT TESTING SLA-D: STUDIES 1 AND 2

Pilots: Participants and Procedures

Pilot testing of the SLA-D iterations (i.e., multiple-choice items) was conducted in two phases: Pilot Study 1 and Pilot Study 2. For both phases of pilot testing, the participants were seventh- and eighth-grade students (12–13 years old) from urban middle schools in the mid-Atlantic region of the United States. Study 1 included 124 participants (75 in seventh grade and 49 in eighth grade) from a single school. Study 2 included 220 participants (170 in seventh grade and 50 in eighth grade) from four other schools.

The SLA-D for Studies 1 and 2 contained 57 and 53 multiple-choice items, respectively. Assessments were completed during the students' scheduled science class periods (blocks); these classes ranged from 50 to 80 minutes in duration. All participants submitted parental consent and student assent forms. Study personnel introduced the study procedure to students, administered the forms, and monitored the testing. All consent procedures and interactions with study subjects were conducted with approval from Montclair State University's Institutional Research Board.

Pilots: Data Analysis

The following analyses were employed to establish validity evidence based on test content, response process, and internal structure. Four key analyses were used to evaluate each item and inform internal structure evidence: (a) percent correct, (b) frequency of responses to each option, (c) discrimination index, and (d) item-total correlation coefficients. We examined each item for the overall percentage of correct responses, expecting that, at most, half of participants would select the correct option for an item. Our participants were not receiving any special preparation in scientific literacy beyond their current science courses in schools. Furthermore, with the goal of developing a measure for use by researchers and educators to assess effects of instruction on scientific literacy, the measure needed to be

TABLE 3
Overview of Research Design

Study →	Pilot Study 1	Pilot Study 2	Prototype Study 3
Research questions	<ol style="list-style-type: none"> How do middle school students respond to items on this? Are indicators of test difficulty (discrimination index, correlations) reasonable for all items? How should the test be revised to better measure scientific literacy? 		<ol style="list-style-type: none"> How do middle school students respond to items on this? Are indicators of test difficulty reasonable for all items? Should the test be revised to better measure scientific literacy? How do participants respond to the attitude and belief items?
Participants	124 seventh- and eighth-grade students four students: think-aloud	220 seventh- and eighth-grade students two students: think-aloud	264 seventh- and eighth-grade students
Procedures	<ul style="list-style-type: none"> SLA-D = 57 MC items Think-aloud with four students Expert review 	<ul style="list-style-type: none"> SLA-D = 53 MC items Two variations Think-aloud with two students 	<ul style="list-style-type: none"> Two versions SLA-D each composed of 26 MC items (11 shared items and 15 unique items) and SLA-MB composed of 25 Likert-type items from 1 to 5 Two ordering variations for each version
Analyses	<ol style="list-style-type: none"> Test score descriptive statistics Item discrimination index (D) Response choice by item Item-total correlation coefficients (r) Thematic analysis 	<ol style="list-style-type: none"> t test to compare test versions Analyses a-e from Study 1 Kuder-Richardson-20 Review and revision of SLA framework 	<ol style="list-style-type: none"> ANOVAs to compare test versions and schools Kuder-Richardson-20 Analyses a-e from Study 1 Principal components analysis SLA-MB Review and revision of SLA framework
Findings	<ul style="list-style-type: none"> D ranged from 0.06 to 0.83, with 19 items demonstrating Ds below 0.30 r ranged from .02 to .59 31 items were revised Seven items were deleted Three new items were created Revised test of 53 multiple-choice items 	<ul style="list-style-type: none"> t-test revealed no difference between versions Ds ranged from 0.05 to 0.82; eight items demonstrated Ds below 0.30 r ranged from .11 to .59 24 items identified for revision 12 items were deleted Revised test of 41 multiple-choice items 	<ul style="list-style-type: none"> ANOVAs revealed no difference between versions Two 26-item versions of final SLA-D Kuder-Richardson-20 = 0.83 for SLA-D1 and 0.82 for SLA-D2 Ds ranged from 0.30 to 0.85 rs ranged from .13 to .62 SLA-MB included three clear factors: Value of science ($\alpha = .80$); science literacy self-efficacy ($\alpha = .72$), and personal epistemology ($\alpha = .88$)

sensitive enough to pick up changes in learning. Thus, items with more than a 50% correct response were either revised to be made more difficult (in Study 1) or were dropped from the measure (in Study 2).

We also examined response frequencies to each item as selected by the participants in each quartile of the distribution of total scores. This method enabled us to revise several item options to make the distractors more attractive, and therefore increase item difficulty. We calculated the discrimination index (D) for each item as the proportion of top total scores (top quartile) who chose the correct response minus the proportion of bottom total scores (bottom quartile) who choose the correct response (Johnson, 1951). The discrimination index provides information as to how well each item discriminates participants from the top and bottom percentiles. Using Hopkins (1998) guidelines for evaluating items based on the D values, we considered items with a D of 0.40 as very strong, 0.30–0.40 as good, and items below 0.30 as needing work (Reynolds, Livingston, & Wilson, 2006).

We calculated the relationship of performance on individual items with the total test score based on the point biserial correlation coefficient. Shaw and Young (2004) offer four recommendations for item retention or deletion based on the item-to-total score correlation coefficients for classroom tests. Specifically they recommend to (1) delete, replace, or revise items with a negative correlation coefficient; (2) replace or rewrite items with zero or (3) low correlation coefficient (i.e., coefficients from .00 to .20, p. 20); and (4) consider using .30 as the “cut-off point for identifying items that may merit retention” indicating that correlations falling in the range of .20–0.30 “. . . are fairly good to quite good items. They could stand as written” (p. 20). It is important to note that these recommendations focus on improving the overall reliability of a teacher’s classroom test used as a summative assessment to evaluate learning that had occurred. Thus, their goal in offering these recommendations was different from our perspective of developing a measure sensitive enough to assess changes in scientific literacy in response to instructional interventions. That is, in this initial development stage we were testing the measure with students who we expected would be relatively naïve to the content, if we followed all of the recommendations we would risk making the test too easy for students with preparation in this content. For that reason, we adhered to the first two recommendations but were more flexible in accepting or retaining items with item-to-total score correlation coefficients lower than .20.

We were also guided by qualitative think-aloud interviews during Studies 1 and 2 from four and two participating students, respectively, to gather response process validity evidence. The classroom teacher was asked to identify students for the think-aloud activity who would be likely to develop rapport and communicate openly with the researcher but who were not necessarily the most accomplished students. These students then completed the test with one of the researchers who prompted each student to “think out loud” while completing the test. The researchers took field notes on a copy of the test, recording comments the students made (including both cognitive and affective responses).

Think-Aloud Interview Insights

The think-aloud interviews provided insights into adolescents’ cognitive as well as affective responses, contributing unique information to the range of evidence the research team considered in deciding to keep, revise, or reject tested items. A specific result of think-aloud feedback involves an item that was developed to assess students’ application of scientific findings to everyday life. In an effort to provide unambiguous directions, the original stem was: “A family decided to make all their decisions based on the results of scientific studies. The adults want the children to get better grades in school and so they set a new rule—all the children must be in bed by 9 PM. Upon which study result was this rule

based?" The think-aloud participants in Study 2 were perplexed by the implausible premise that a family would make all decisions based on the results of scientific studies, although they were able to understand the question. Based on their feedback, we revised the stem to a more relatable scenario: "Arturo's parents want him to get better grades in school. His mother read a research study on the topic. After reading the study, she decided that from now on Arturo needed to be in bed by 9 PM. Which of these studies did Arturo's mother read?"

A related insight, based on observations of students' affective responses during the think-aloud interviews, was that they were energized by questions that they found to be inherently worthwhile or "fun" based on the topic or context. Consistent with findings of Drennan (2003), the qualitative information generated by the think-aloud interviews helped inform decisions about which items to eliminate, and bolstered the research team's understanding of students' stamina and willingness to work through complex questions.

Pilot Study 1 Results

The overall mean number correct score for this test was 32 of 57 (56%) with minimum and maximum scores of 14 and 51 (24% and 89%), and the distribution resembled a normal curve. The discrimination index (D) ranged from 0.06 to 0.83 with 38 items demonstrating D s of 0.30 and greater; this indicated how well the items discriminated low scorers from top performers on the test overall. The bi-serial correlations between individual items and the total score ranged from .02 to .59, with a median of 0.32. As noted above, Shaw and Young (2004) recommend revision or deletion of items with correlations coefficients less than .20.

We gathered additional evidence of validity of test content by sharing a copy of the Pilot Study 1 test with a group of 25 other science education researchers with current Science Education Partnership Awards (SEPA) from the National Institutes of Health. They reviewed and commented on items during a session at a SEPA Principal Investigators' annual meeting. A member of our research team gathered their feedback and reported it to our team.

Items with correlation coefficients lower than .20 and D s lower than 0.30 were closely scrutinized and then deleted or revised based on a number of criteria including the distribution of responses to each distractor, the overall number of items aligned with the five target components of scientific literacy, feedback on items from SEPA colleagues (particularly regarding comments about cultural appropriateness), and responses from the think-aloud interviews. Through detailed discussions of the above data for each item, we decided to revise 31 items, delete seven items, and create three new items for the next pilot test (Study 2).

Pilot Study 2 Results

We created two orderings of the 53-item SLA-D to address any possible order or fatigue effects on students' responses to the items. A series of analyses of variances (ANOVAs) indicated that the ordering did not lead to significant differences for the total score; thus, for the following analyses, responses from both versions were considered together.

Based on the 220 observations, the Kuder–Richardson equation 20 reliability for the 53 items was 0.90, suggesting that responses to the multiple-choice items were collectively reliable. Because responses were scored on a binary scale (i.e., correct or incorrect), we used the Kuder–Richardson reliability equation as equivalent to the Cronbach alpha. Although we considered the items on the SLA-D to reflect a single construct of demonstrated scientific

literacy, we were also interested in how items related to each of the five components of our framework (see below for an explanation of our framework reduction from six to five components). The reliability for the individual components was lower, ranging from 0.46 to 0.76, with the magnitude of the coefficient approximately proportional to the square root of the number of items in the component. It is important to note that because dividing up items across the components allowed fewer items per component (which decreases reliability) and because the subjects were expected to be naïve to the material, the relatively low reliability scores on the individual components were expected. However, we interpret the substantial decline in reliability when examining the five individual components compared with the full complement of items as providing initial evidence regarding the internal structure of the SLA-D, suggesting that it is most appropriately viewed as a single construct.

We followed the same procedures for item review as in Pilot Study 1 to identify items for revision or deletion. As before, our decisions concerning item revisions and deletions were guided by our goals of aligning with our framework of scientific literacy as well as the statistical results (detailed below), and two additional think-aloud interviews. The overall mean number correct score for this test was 26.6 of 53 (50%), slightly lower than in Study 1, with a wider range of scores from 4 to 52 (8–98%). Overall, 31 of the 53 items were correctly answered by 50% or fewer students. Items to which more than 50% of participants responded correctly were considered “too easy” and were marked for either deletion or revision to ensure an adequate sampling of the construct of interest (AERA, APA, NCME, 1999).

The *Ds* ranged from 0.05 to 0.82, with only eight of the 53 items demonstrating *Ds* of 0.29 or lower (an improvement over Study 1). Biserial correlations between the item score and the total score ranged from .11 to .59. Based on our discussion and analysis of each item, we deleted 12 items and identified 24 items for revisions, primarily for simplification of stems and clarification of language. The resulting multiple-choice measure comprised 41 items.

In addition to clarifying items, the team also used the Study 2 results to review and adjust the overall framework for demonstrated scientific literacy underpinning the measure development. We did so by mapping backwards from constructed items to the initial components and considering the reliability analysis for each component. Based on this, we felt that combining what was initially conceived as two separate components (“science media literacy” and “science and society”) into a single component provided an adequate and more parsimonious framework. Although as stated earlier, we believe that the SLA-D is most appropriately viewed as a single construct. In Table 4, we illustrate the alignment of sample items with each component of the revised conceptual framework.

PROTOTYPE TESTING

The goal of Prototype Study 3 was to evaluate our prototype measure and identify any final items for revision or deletion (see Table 3 for the research questions). Based on our findings from the pilot studies, in Study 3 we tested four variations of the SLA that included both the SLA-D and SLA-MB subscales (the latter was not tested in Studies 1 and 2).

Measures

SLA-D. The SLA-D included two versions (i.e., SLA-D1 and SLA-D2) each composed of 26 items. Eleven items are shared between the two versions and each version includes 15 unique items. The two versions of the SLA-D represent 41 unique items in all. The

TABLE 4
Final Scientific Literacy Framework and Sample Items

Component	Sample Item
Role of science <ul style="list-style-type: none">Identify questions that can be answered through scientific investigation;Understand the nature of scientific endeavors;Understand generic science terms/concepts.	<p>A country has a high number of decayed teeth (cavities) per person. Which question about tooth decay can only be answered with scientific experiments?</p> <p>(a) Do the men in this country have more tooth decay than women? (b) Would putting vitamin D in the water supply affect tooth decay? (c) Has the number of decayed teeth increased in the past 10 years? (d) Is tooth decay more common in some parts of the country than others?</p>
Scientific thinking and doing <i>observational and analytical abilities</i> <ul style="list-style-type: none">Describe natural phenomena;Recognize patterns;Identify study variables;Ask critical questions about study design;Reach/evaluate conclusions based on evidence.	<p>The principal of Riley middle school wants to remove candy and soda (pop) vending machines. In their place, she wants to put in healthy food machines. She wants to know what her students will think of these changes. What would be the best way to get an accurate answer to this question?</p> <p>(a) Give a survey to all students who play on sports teams. (b) Give a survey to all students who attend a health fair. (c) Give a survey to every 20th student on a list of all students. (d) Give a survey to all students who use the vending machines.</p>
Science and society <i>Critique of scientific findings described in the popular media</i> <ul style="list-style-type: none">Apply scientific conclusions to daily life;Understand the role of science in policy decision-making;Develop questions to assess validity of scientific reports;Question the sources of science reporting;Identify scientific issues underlying policy decisions.	<p>A student finds a website created by the “No Homework Committee.” He wants to find out the reasons for and against assigning homework to students. Is this a trustworthy source of information?</p> <p>(a) Yes. This group is against homework and knows all of the arguments. (b) Yes. Information on Web sites is always balanced and correct. (c) No. This group might give more attention to arguments against homework. (d) No. This group is probably not very good at arguing for or against homework.</p>

a

(Continued)

TABLE 4
Continued

Component	Sample Item																
Mathematics in science <ul style="list-style-type: none">• Use mathematics in science;■ Understand the application of mathematics in science.	<p>What percent of the sample of people shown in the graph is older than 15 years?</p> <p>(a) 20% (b) 30% (c) 40% (d) 50%</p> <table border="1"><caption>Data for Sample Item Graph</caption><thead><tr><th>Age (Years)</th><th>Percent (%)</th></tr></thead><tbody><tr><td>12</td><td>10</td></tr><tr><td>13</td><td>10</td></tr><tr><td>14</td><td>10</td></tr><tr><td>15</td><td>20</td></tr><tr><td>16</td><td>10</td></tr><tr><td>17</td><td>10</td></tr><tr><td>18</td><td>10</td></tr></tbody></table>	Age (Years)	Percent (%)	12	10	13	10	14	10	15	20	16	10	17	10	18	10
Age (Years)	Percent (%)																
12	10																
13	10																
14	10																
15	20																
16	10																
17	10																
18	10																
Motivation and beliefs <ul style="list-style-type: none">■ Value of science (Wigfield & Eccles, 2000);• Self-efficacy for scientific literacy (adapted from Kettlehut, 2010);■ Source and certainty of scientific knowledge (Conley et al., 2004).	<p>Value: In general, I find working on science assignments (1: very boring to 5: very interesting)</p> <p>Self-efficacy: I can use science to make decisions about my daily life (1: strongly disagree to 5: strongly agree)</p> <p>Personal epistemology: All questions in science have one right answer (1: strongly disagree to 5: strongly agree)</p>																

SLA-D1 and SLA-D2 can be found in Appendices 1 and 2 in the Supporting Information, respectively.

SLA-MB. The SLA-MB is composed of three motivation and belief scales assessed on a 1–5 scale: value of science (six items), self-efficacy for scientific literacy (eight items), and personal epistemology of science (11 items—reverse coded; see Appendix 3 in the Supporting Information for these scales). Higher scores indicate stronger value, self-efficacy, and more sophisticated beliefs about science. We adapted Wigfield and Eccles (2000) measure of achievement value to assess students' value of science by replacing “math” with “science” in each of the six items. Scientific literacy self-efficacy was assessed with eight items. Four of these items came from Kettlehut's (2010) measure and we drafted four new items to better align with our conceptualization of scientific literacy. Specifically, we added items to assess student's confidence reflective of topics in the role of science, science and society, science media literacy, and mathematics in science. Beliefs about the source and certainty of knowledge in science were assessed with two scales from Conely et al.'s (2004) measure of personal epistemology in science. These items were worded such that low scores demonstrated more sophisticated epistemological beliefs, and as recommended by Conley et al. (2004), we reverse coded these items in our analyses to align the scores with the other two scales. The SLA-MB can be founded in Appendix 3 in the Supporting Information.

Variations. During the prototype testing in addition to testing two SLA-D versions, we also varied the *order* in which participants received the SLA-D version with the SLA-MB scales. This resulted in four variations of the SLA that were tested: (1) SLA-D1 → SLA-MB; (2) SLA-MB → SLA-D1; (3) SLA-D2 → SLA-MB; (4) SLA-MB → SLA-D2. We engaged in this variation to ameliorate any order effects in completing the assessments.

Participants and Procedures

Data collection was initiated with 321 middle school students in seventh or eighth grade from five schools in Northern New Jersey. Data collection took place in May and June 2012 during students' scheduled science classes. All participants submitted parental consent and student assent forms and all study activities were conducted with approval from Montclair State University's Institutional Research Board. The enrollment response rate was 26% overall, with a range by school of 17–45%, a likely reflection of the written parental consent required to participate. Female students were slightly overrepresented (56%) and the participants reported a range of ethnicities including Hispanic (34%), White (24%), Asian (21%), African American (14%), and other or multiple (7%).

Evaluation of the Prototype

Consistency Across Schools, Grade Levels, and Variations of SLA-D. We deleted responses from four students who had two or more missing responses on the SLA-D portion, leaving data from 317 participants for analysis. To determine if there were any significant differences related to the schools, grade levels, or variations of the SLA, we conducted a univariate ANOVA. School, grade, measure variation, and their two and three-factor interactions served as the independent variables, with the score on the SLA-D items as the dependent variable. Because of the higher order interactions, we were not able to test for the homogeneity of variance assumption, but the Shapiro—Wilk (1965) test for normality

TABLE 5
Description of Participants in Study 3

Characteristic	Number of Participants		
	Seventh	Eighth	All
School by grade			
1	0	38	38
2	0	133	133
3	41	0	41
4	20	32	52
Total	61	203	264
School by gender	Male	Female	All
1	13	25	38
2	65	68	133
3	17	24	41
4	25	27	52
Total	120	144	264
Ethnicity by gender	Male	Female	All
Asian	29	35	64
Black or African American	9	14	23
Hispanic or Latino	31	52	83
White	41	34	75
Other ethnicities (American Indian, Pacific Islander); or two or more ethnic backgrounds indicated; or missing	10	9	19
Total	120	144	264

of the residuals indicated the residuals were normally distributed and there were no outliers ($p < .23$), which gave us confidence that we had not violated either the homogeneity of variance or the normality of residuals assumptions. ANOVA results indicated a significant difference across schools [$F(4, 289) = 38.67; p < .0001, \eta^2 = 0.35$, where η^2 is the effect size] but no significant differences across test variations [$F(3, 289) = 2.01; p < .11, \eta^2 = 0.02$] or grade level [$F(1, 289) = 2.16; p < .14, \eta^2 = 0.007$], nor in the 2 two- and three-factor interactions. Follow-up analyses used Duncan's multiple range test (Duncan, 1975), a multiple comparison test that tests pairs of group means while considering the multiple comparisons problem associated with a test like the multiple t -test. These tests indicated that School 2 had statistically significantly higher scores than all others ($M = 17.1, SD = 4.9$), Schools 1 and 3 were similar to each other and different from all others ($M_s = 14.8, 13.4, SD_s = 5.0$), School 4 ($M = 11.1, SD = 4.0$) and School 5 ($M = 7.7, SD = 2.5$) were different from all other schools and scored the lowest in the order presented. In addition, the students in School 5 were administered the measure on the last day of school. Based on anecdotal evidence from the researcher administering the SLA at this school and the combined teaching experience of other members of the research team, we determined that the extremely low scores from both grades in School 5 could well be due to this contextual variable. Therefore, we dropped all data from School 5 for all remaining analyses ($n = 54$, of which one had been dropped because of too many missing responses), leaving data from 264 participants for further analysis. Table 5 provides complete demographic information for the participants used in the remaining analyses.

SLA-D Evaluation. For the 264 responses from the four schools retained for the remainder of our investigation, the mean SLA-D was 15 correct of 26 (58%), with a range of 2–25 (8–96%). An ANOVA was performed in which school, grade, measure form, and their two- and three-factor interactions served as the independent variables and score on the multiple-choice items as the dependent variable. The Shapiro–Wilk (1965) test for normality of the residuals indicated that the residuals were normally distributed and there were no outliers ($p < .15$), which, again, gave us confidence that we had not violated either the homogeneity of variance or the normality of residuals assumptions. As in the previous analyses, the ANOVA results indicated a significant difference across schools [$F(3, 244) = 13.93$; $p < .0001$, $\eta^2 = 0.14$]. Follow-up univariate analyses indicated that School 4 ($M = 11.0$, $SD = 4.0$) had a significantly lower mean score than the other schools on the SLA-D. This was indicated in our initial comparison across the five schools. We elected to keep School 4 in our analyses because the overall population in School 4 was the most diverse with respect to ethnicity and socioeconomic status and we wanted to ensure that decisions we made about this measure would reflect a wide range of student groups. The goal of this project is to create a measure of scientific literacy that would be sensitive enough to pick up variability across groups and educational experiences. If we dropped school 4 from our analyses, we may have elected to delete more items and consequently develop a test that was too difficult or culturally limited to assess a wide variation in demonstrated scientific literacy among diverse students.

There were no statistically significant differences in the four test variations [$F(3, 244) = 2.64$; $p < .06$, $\eta^2 = 0.03$], or grade level [$F(1, 244) = 2.21$; $p < .14$, $\eta^2 = 0.01$]. Since some schools provided test results for only 1 grade, it was not possible to test for the interactions with school and grade, but none of the calculated higher order interactions were statistically significant. The lack of differences on these measures suggests that the order of the measure in terms of multiple-choice items or motivation and belief scales coming first or second had no bearing on participants' scores on the multiple-choice items, and that the two versions (SLA-D1 and SLA-D2) are not statistically significantly different (i.e., they are equivalent). The means and standard deviations of scores on the multiple-choice items by school, grade level, and test form can be found in Table 6.

The Kuder–Richardson equation 20 reliability (1937) for the two versions of the SLA-D was 0.83 and 0.82, respectively. The discrimination indices for the 41 items that make up both versions of the SLA-D ranged from 0.30 to 0.85 with a mean and median of 0.58. The highest percent correct for any one item was 89% and the lowest was 22%. Overall 12 of the 41 items (29%) were selected correctly by 50% or fewer students.

To assess the reasonableness of the scientific literacy components assessed by the SLA-D items, we conducted a principal components factor analysis. The purpose was to test whether the items aligned with each component could be used as independent measures, despite our previous finding of low reliability by component (see Study 2). The analysis indicated that the SLA-D for this participant pool also assessed a single factor of demonstrated scientific literacy. As such we recommend using the SLA-D as a single measure of demonstrated scientific literacy.

SLA-MB Evaluation. One of the 264 participants did not respond to the motivation and belief scale items, leaving 263 responses for analysis on SLA-MB scales. Exploratory principal components factor analysis on the 25 items of the SLA-MB indicated that they formed three well-defined and unique components with the items having high eigenvalues (minimum eigenvalues of 0.62, 0.55, and 0.36 on the first three components). This suggests that the SLA-MB assesses three distinct sets of beliefs associated with scientific literacy

TABLE 6
Study 3: Prototype SLA-D Mean Scores by Version, Variation, School, and Grade

Test Form	Grade → n	School					
		1		2		3	
		Eighth		Eighth		Seventh	
						Seventh	Eighth
1: SLA-D1 → SLA-MB	M (SD)	9 17.1 (4.2)	35 18.3 (4.7)	13 13.8 (4.9)	4 11.5 (2.1)	9 13.4 (3.5)	70 16.3 (4.9)
2: SLA-MB → SLA-D1	M (SD)	8 13.6 (4.1)	33 18.1 (5.1)	7 14.1 (5.5)	5 10.6 (5.9)	8 13.0 (3.3)	61 15.8 (5.4)
3: SLA-D2 → SLA-MB	M (SD)	10 13.1 (6.8)	32 15.6 (4.9)	10 12.9 (5.5)	5 8.0 (2.4)	7 10.6 (4.3)	64 13.6 (5.5)
4: SLA-MB → SLA-D2	M (SD)	11 15.3 (3.9)	33 16.5 (4.7)	11 12.9 (4.9)	6 9.3 (3.9)	8 10.5 (4.0)	69 14.4 (5.1)
All	M (SD)	38 14.8 (5.0)	133 17.1 (4.9)	41 13.4 (5.0)	20 9.8 (3.9)	32 12.0 (3.8)	264 15.0 (5.3)

TABLE 7
Study 3: Descriptive Statistics and Reliability Scores for the SLA-MB
(*n* = 263)

Component of SLA-MB	<i>M</i>	<i>SD</i>	<i>A</i>
Value of science (Wigfield & Eccles, 2000)	3.9	0.7	0.80
Self-efficacy for scientific literacy (adapted from Kettlehut, 2010)	3.8	0.6	0.72
Source and certainty of scientific knowledge (Conley et al., 2004)	3.7 ^a	0.8	0.88

Minimum score = 1, Maximum score = 5.

^aReverse coded.

TABLE 8
Study 3: Correlation Matrix for SLA-MB and SLA-D

Study 3: Correlation Matrix for SLA-MB and SLA-D (<i>N</i> = 263)				
Component of SLA-MB	Value	Self-Efficacy	Knowledge ^a	SLA-D
Value of science (Wigfield & Eccles, 2000)	1.000	.530 (<i>p</i> < .0001)	−.110 (<i>p</i> = .060)	.100 (<i>p</i> = .120)
Self-efficacy for scientific literacy (adapted from Kettlehut, 2010)		1.000	.050 (<i>p</i> = .400)	.400 (<i>p</i> < .0001)
Scientific knowledge is uncertain and constructed ^a (Conley et al., 2004)			1.000	.370 (<i>p</i> < .0001)
SLA-D				1.000

^aReverse coded.

and that these scales can be used independently of one another. Each scale consistent with previous findings demonstrated sound reliability (value of science: $\alpha = .80$; self-efficacy for science literacy: $\alpha = .72$; source and certainty of scientific knowledge: $\alpha = .88$). Table 7 provides the means, standard deviations, and Cronbach’s alpha for each of these components.

SLA-MB Relation to SLA-D Scores

Correlations among the three SLA-MB components and the SLA-D are shown in Table 8. Among the three constructs, there was a strong positive correlation ($r = .53$) between the mean score on the value of science and self-efficacy. This is consistent with what is seen in the expectancy-value research, that students frequently report valuing tasks that they feel confident in achieving (or the reverse). There was no correlation between the students’ total score on the SLA-D and the mean response on the value of science construct, and moderate correlation with self-efficacy ($r = .40$) and personal epistemology ($r = .37$). The relation between self-efficacy and achievement suggests that students may have a good sense for their ability to engage in scientific literacy. Furthermore, the moderate correlation with personal epistemology demonstrates a relation between understanding knowledge as tentative and constructed with students’ ability to demonstrate scientific literacy. Visual inspection of scatter plots (not shown) confirmed these interpretations.

LIMITATIONS

Our validation argument is limited. First, we do not provide evidence based on relations to other variables. This first omission is a serious limitation of our current work. Future research with this measure needs to establish relations between scores on the SLA-D with other criteria that provide evidence of scientific literacy, such as science grades, performance evaluations of scientific literacy, or scores on a similar test. Unfortunately, the development process at the time of this investigation did not allow for such comparisons as data were gathered from students anonymously. Furthermore, while the measure was still in development we did not think that this would be an adequate use of our resources. This is a much needed next step to add to the validation argument for the SLA. The SLA should be tested as a pre–post assessment tool for lessons, units, or courses in which scientific literacy is systematically addressed. Gains in scores at posttest, especially relative to an independent assessment of student learning, would provide further evidence that the SLA assesses scientific literacy. Future work needs to establish this evidence to support widespread use.

Second, our participant pool for the prototype study was limited to four schools in one district, in one U.S. state. This limits the generalizability of the findings in this investigation. Further testing of this SLA needs to be conducted with participants in other states and countries to ensure the effectiveness of this tool in assessing scientific literacy in different cultural contexts. It should be noted that our participants represented a range of ethnicities and socioeconomics levels.

Third, additional content and response process evidence could be garnered from a sample of practicing middle school level teachers. While initial versions of the SLA-D were shared with four teachers (two middle level and two high school) and the development team included a former middle school science teacher and a science educator who develops programming for middle school students, a review of these tools by a larger group of teachers could prove informative in moving forward with revised versions of this measure.

Fourth, we do not provide any evidence based on test consequences. Evidence based on consequences of testing should demonstrate that any negatives associated with taking a test are outweighed by positive outcomes; furthermore, evidence of this type should demonstrate the likelihood of intended benefits actually occurring (AERA, APA, NCME, 1999). The inclusion of testing consequences as a source of validity is described as the “most contested validity territory” (Cizek, Rosenbertg, & Koons, 2008, p. 398). Some scholars (e.g., Kane, 2001; Linn, 1997; Messick, 1995; Shepard, 1997) argue for including this in conceptions of validity because “negative consequences can render a score’s use as unacceptable” (Kane, 2013, p. 1). However, others (e.g., Borsboom, Mellenbergh, & van Heerden, 2004; Cizek et al., 2008; Cizek, 2012; Dwyer, 2000; Popham, 1997) have argued against its inclusion in validity theory for reasons such as “the social consequences of score use do not bear on the validity of score interpretations” (Cizek, 2012, p. 3).

Fifth, we have pragmatic concerns about the length of the SLA-D and the feasibility of students’ completing this measure in one 40-minute class session. To address this concern, we recommend shortening the SLA-D to 19 items (2 minutes per item). Following the same iterative analyses and processes described in the pilot studies as well as contextual factors such as amount of reading required for an item (we took on a “less is better” approach) and an adequate distribution of items across components of our scientific literacy framework, we reviewed the 41 items of the SLA-D and identified seven items to delete from each version of the SLA-D. If these deletions are made, our final recommended prototype measure includes two versions of the SLA-D, each with 19 multiple-choice items (nine shared items

between versions for a total of 29 unique items) and all 25 items from the motivation and belief scales. The statistical analyses on the items from these reduced SLA-D versions based on the 264 responses to the 26-item tests indicated results very similar to the results presented in our Prototype test (data not shown). Therefore, we feel that it is possible to make reasonable inferences about middle schools students' scientific literacy based on the reduced measure. In Appendices 1 and 2 in the Supporting Information, we indicate which items to redact from the SLA-D versions.

DISCUSSION OF THE SLA

The SLA is intended to assess middle schools students' sense of field/discipline general scientific literacy. The SLA is designed to be administered in one class period (40–50 minutes) via a paper and pencil format. The SLA has two parts: the SLA-D that assesses five components of demonstrated scientific literacy and the SLA-MB modified from three existing scales that measure scientific literacy motivation and beliefs. There are two versions of the SLA-D portion. Each includes 26 multiple-choice items (11 shared and 15 unique items on each version), presented in Appendices 1 and 2 in the Supporting Information, including directions for shortening the measure if needed. The SLA-D items are written at, or below, the sixth-grade level according to the Flesch–Kincaid index. The SLA-MB is composed of three subscales for a total of 25 Likert items that include the three motivation and belief scales (Appendix 3 in the Supporting Information).

Our findings support using the SLA-D and SLA-MB to assess middle school students' scientific literacy. Through careful attention to the standards for developing validity arguments (AERA, APA, NCME, 1999), we have provided comparative validity evidence related to test content, response process, and internal structure. The results of our iterative process of item construction, administration, and revision provide support that the SLA-D and SLA-MB align with the underlying conceptualization of scientific literacy that we sought to assess. In addition, the development of this measure was guided by experts from SEPA, an interdisciplinary research team, and a sound conceptualization of scientific literacy based on the extant literature.

The SLA-D items in both versions demonstrate good reliability, and the items on each adhere to recommended guidelines for percent correct, discrimination index, item-total correlation coefficients, and frequency distribution of distractors selected, all of which provide evidence for the strong internal structure of this measure. Furthermore, the lack of statistical or practical difference in scores from students responding to the two versions of the SLA-D suggests that these versions are assessing equivalent information. For these reasons, we recommend the use of this measure with middle school students and encourage users to evaluate the reliability in their data and consider the appropriateness of this tool for providing valid evaluations of scientific literacy in the contexts in which it is used.

The correlations among the SLA-D score and the SLA-MB scales further inform our understanding of scientific literacy as abilities, motivation, and beliefs. The interesting distinctions in correlations among value of science, self-efficacy for scientific literacy, and personal epistemology indicate a potentially developing sense of scientific literacy in these students. For instance, students who value science would likely feel that they are good at science, but their personal epistemology may not yet be well formed so it is unrelated to appreciation and ability. The low correlation between total item score and the value of science score reflects the possible disconnect between appreciation and ability; the higher correlation between total item score and self-efficacy reflects the link of self-assessment and external assessment. The moderate correlation between total item score and the personal

epistemology score is an indication that those who understand the nuance of science will also be better at science as measured by an external evaluation.

IMPLICATIONS AND CONCLUSIONS

While the current tool is still a work in progress, we see implications for the present work to inform both pedagogical practice and theoretical development of the construct of scientific literacy. Pedagogically, the relations among the SLA-D and SLA-MB illustrate the importance of teaching not just the content of science literacy but also the need to allow students opportunities to develop beliefs and values that support the use of science in their lives. Classroom teachers could use the SLA as a formative assessment at the beginning of the academic year to target aspects of scientific literacy (knowledge, beliefs, and values) for instruction during the school year. Furthermore, teachers could use the SLA-MB to begin a conversation with their students about the students' personal epistemology, value, and motivation for science. Exposing such beliefs could be a first step in helping students to better understand themselves in relation to science.

Theoretically, the SLA provides a measure of demonstrated knowledge as well as students' beliefs and motivation. To achieve the goal of a scientifically literate society, individuals need to be more than knowledgeable of the science content, they must also value that content and be open to it as a source of information for decision making. The correlational results presented here, while still tentative, indicate that a relationship between demonstrated knowledge and motivation and beliefs exist. Furthermore, we identified three key areas of motivation and belief for inclusion in conceptions of scientific literacy: personal epistemology, self-efficacy, and value. The field of motivation offers a variety of other constructs that may also prove informative. Thus, these findings tentatively suggest that further theoretical and empirical investigation into the nature of knowledge, motivation, and a belief as part of scientific literacy is warranted.

The concept of scientific literacy "has become an internationally well-recognized educational slogan, buzzword, catchphrase, and contemporary educational goal" (Laugksch, 2000, p. 71) despite the lack of agreement on just what it *is* (see Dillon, 2009; Holbrook & Rannikemae, 2009; Laugksch, 2000; Roberts, 2007). We have developed a measure of scientific literacy that is appropriate for middle school students. It is not designed to assess specific content knowledge, such as Newton's law of gravity or Boyle's law of thermodynamics, but measures a functional understanding and appreciation of science.

While the SLA has met standard measures of internal structure and reliability, we do not consider the test to be fully validated in either the technical or the vernacular sense. We see the studies presented here as the intermediate steps of a work in progress and would like interested groups to use and evaluate this test to develop a wide group validation along the lines of *crowd sourcing*. We hope that through this mechanism a reasonable and useful test of scientific literacy can be fully developed from our work. We believe that such a tool is necessary to the promotion of scientific literacy that in turn can aid in combating ignorance about the importance of science and promote rational scientific policy decision making in a democratic society.

The authors would like to thank Mark Kaelin under whose guidance the project was conceived and the team assembled, and Tony Beck of the SEPA program for his encouragement and support. We appreciate Lisa Abrams, Mike Kennedy, Marian Passannante, Kristin Bass, and Ron Vangi for their expert advice, and Doug Larking for reading a prior version of this paper. Finally, we thank the middle school science teachers and students who good-naturedly participated in our testing and understood the importance of their contribution to the research.

REFERENCES

- AAAS Project 2061 Science Assessment Website. (2011). American Association for the Advancement of Science, Project 2061. Retrieved April 9, 2011, from <http://assessment.aaas.org>
- Aikenhead, G. (2011). Towards a cultural view on quality science teaching. In D. Corrigan, J. Dillon, & R. Gunstone (Eds.), *The professional knowledge based of science teaching* (pp. 107–127). New York: Springer.
- Alexander, P. A., Kulikowich, J. M., & Schulze, S. K. (1994). How subject-matter knowledge affects recall and interest. *American Educational Research Journal*, 31(2), 313–337.
- American Association for the Advancement of Science. (1989). *Project 2061: Science for all Americans*. Washington, DC: Author. Retrieved February 12, 2011, from <http://www.project2061.org/publications/sfaa/online/sfaatoc.htm>.
- American Association for the Advancement of Science. (1993). *Benchmarks for scientific literacy*. Project 2061. New York: Oxford University Press.
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Arons, A. B. (1983). Achieving wider scientific literacy. *Daedalus*, 112(2), 91–122.
- Ausubel, D. P. (1977). The facilitation of meaningful verbal learning in the classroom. *Educational Psychologist*, 12(2), 162–178.
- Baldwin, R. S., Peleg-Bruckner, Z., & McClintock, A. H. (1985). Effects of topic interest and prior knowledge on reading comprehension. *Reading Research Quarterly*, 20(4), 497–504.
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.
- Berry, A., Loughran, J., & Mulhall, P. (2007). Values associated with representing science teachers' pedagogical content knowledge. In D. Corrigan, J. Dillon, & R. Gunstone (Eds.), *The re-emergence of values in science education* (pp. 149–163). Rotterdam, The Netherlands: Sense Publishers.
- Bøe, M. V. (2012). Science choices in Norwegian upper secondary schools: What matters? *Science Education*, 96, 1–20.
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111, 1061–1071.
- Braten, I., Stromso, H., & Samuelson, M. S. (2008). Are sophisticated students always better? The role of topic-specific personal epistemology in the understanding of multiple expository texts. *Contemporary Educational Psychology*, 33, 814–840.
- Britner, S. L., & Pajares, F. (2001). Self-efficacy beliefs, motivation, race, and gender in middle school science. *Journal of Women and Minorities in Science and Engineering*, 7, 271–285.
- Bryan, R. R., Glynn, S. M., & Kittleson, J. M. (2011). Motivation, achievement, and advanced placement intent of high school students learning science. *Science Education*, 95(6), 1049–1065.
- Business Higher Education Forum (BHEF). (2011). *Creating the workforce of the future: The STEM interest and proficiency challenge* (2011). Retrieved August 31, 2012, from http://www.bhef.com/publications/documents/BHEF_Research_Brief-STEM_Interest_and_Proficiency.pdf.
- Bybee, R. W. (2008). Scientific literacy, environmental issues, and PISA 2006: The 2008 Paul F-Brandwein lecture. *Journal of Science Education and Technology*, 17, 566–585.
- Chen, J. A., & Pajares, F. (2010). Implicit theories of ability of Grade 6 science students: Relation to epistemological beliefs and academic motivation and achievement in science. *Contemporary Educational Psychology*, 35(1), 75–87.
- Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods*, 17(1), 31–43.
- Cizek, G. J., Rosenbergt, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement*, 68, 398–412.
- Conley, A. M., Pintrich, P., Vekiri, I., & Harrison, D. (2004). Changes in epistemological beliefs in elementary science students. *Contemporary Educational Psychology*, 29, 186–204.
- Cross, R. T., & Price, R. F. (1992). *Teaching science for social responsibility*. Sydney, Australia: St Louis Press.
- DeBoer, G. E. (2000). Scientific literacy: Another look at its historical and contemporary meanings and its relationship to science education reform. *Journal of Research in Science Teaching*, 37, 582–301.
- Dillashaw, F. G., & Okey, J. R. (1980). Test of the integrated science process skills for secondary science students. *Science Education*, 64(5), 601–608.
- Dillon, J. (2009). On scientific literacy and curriculum reform. *International Journal of Environmental & Science Education*, 4, 201–213.
- Drennan, J. (2003). Cognitive interviewing: Verbal data in the design and pretesting of questionnaires. *Journal of Advanced Nursing*, 42(1), 57–63.

- Duit, R., & Treagust, D. F. (2003). Conceptual change: A powerful framework for improving science teaching and learning. *International Journal of Science Education*, 25, 671–688.
- Duncan, D. B. (1975). t-tests and intervals for comparisons suggested by the data. *Biometrics*, 31, 339–359.
- Dwyer, C. A. (2000). Excerpt from validity: Theory into practice. *The Score*, 22, 6–7.
- Eccles, J., Barber, B., & Jozefowicz, D. (1999). Linking gender to educational, occupational, and recreational choices: Applying the Eccles et al. model of achievement-related choices. In W. B. Swann, Jr., J. H. Langlois, & L. A. Gilbert (Eds.), *Sexism and stereotypes in modern society: The gender science of Janet Taylor Spence* (pp. 153–192). Washington, DC: American Psychological Association.
- Eccles, J., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, 53(1), 109–132.
- Fuchs, B. A. (2008). Teaching Science as Inquiry: Successes and Challenges in the U.S. Presented at NIH Blueprint K-12 Neuroscience Research–K-12 Education Workshop. Rockville, MD.
- Gauld, C. (1982). The scientific attitude and science education: A critical reappraisal. *Science Education*, 66, 109–121.
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15, 309–334.
- Hamm, M. (1992). Achieving scientific literacy through a curriculum connected with mathematics and technology. *School Science and Mathematics*, 92, 6–9.
- Hazen, R. M., & Trefil, J. (1991). *Science matters: Achieving scientific literacy*. New York: Doubleday.
- Hodson, D. (1999). Going beyond cultural pluralism: science education for socio-political action. *Science Education*, 83, 775–796.
- Hofer, B. K. (2000). Dimensionality and disciplinary differences in personal epistemology. *Contemporary Educational Psychology*, 25, 378–405.
- Hofer, B. K., & Pintrich, P. R. (1997). The development of epistemological theories: Beliefs about knowledge and knowing and their relation to learning. *Review of Educational Research*, 67, 88–140.
- Holbrook, J., & Rannikemae, M. (2007). The nature of science education for enhancing scientific literacy. *International Journal of Science Education*, 29, 1347–1362.
- Holbrook, J., & Rannikemae, M. (2009). The meaning of scientific literacy. *International Journal of Environmental & Science Education*, 4(3), 275–288.
- Hopkins, K. D. (1998). *Educational and psychological measurement and evaluation* (8th ed.). Boston: Allyn and Bacon.
- Jarman, R., & McClune, B. (2007). *Developing scientific literacy: Using news media in the classroom*. Maidenhead, England: McGraw-Hill International.
- Jenkins, E. (2003). School science: Too much, too little, or a problem with science itself? *Canadian Journal of Science, Mathematics and Technology Education*, 3(2), 269–274.
- Johnson, O. K. (1951). The effect of classroom training up on listening comprehension. *Journal of Communication*, 1, 58.
- Kane, M. T. (1994). Validating interpretative arguments for licensure and certification examinations. *Evaluation & the Health Professions*, 17, 133–159.
- Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38, 319–342.
- Kane, M. T. (2012). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50, 1–73.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73.
- Ketelhut, D. J. (2010). Assessing gaming, computer and scientific inquiry self-efficacy in a virtual environment. In L. A. Annetta & S. Bronack (Eds.), *Serious educational game assessment: Practical methods and models for educational games, simulations and virtual worlds*. Amsterdam, The Netherlands: Sense Publishers.
- Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2(3), 151–160.
- Lambert, J. (2006). High school marine science and scientific literacy: The promise of an integrated science course. *International Journal of Science Education*, 28, 633–654.
- Laugksch, R. C. (2000). Scientific literacy: A conceptual overview. *Science Education*, 84(1), 71–94.
- Laugksch, R. C., & Spargo, P. E. (1996). Construction of a paper-and-pencil *Test of Basic Scientific Literacy* based on selected literacy goals recommended by the American Association for the Advancement of Science. *Public Understanding of Science*, 5(4), 331–359.
- Lent, R. W., Brown, S. D., & Gore, P. A., Jr. (1997). Discriminant and predictive validity of academic self-concept, academic self-efficacy, and mathematics-specific self-efficacy. *Journal of Counseling Psychology*, 44, 307–315.

- Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice*, 16, 14–16.
- Liu, X. (2009). Beyond science literacy: Science and the public. *International Journal of Environmental & Science Education*, 4(3), 301–311.
- Messick S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741–749.
- Millar, R., & Osborne, J. (1998). Beyond 2000: Science education for the future. The report of a seminar series funded by the Nuffield Foundation. London: School of Education, King's College London.
- Miller, J. D. (1983). Scientific literacy: A conceptual and empirical review. *Daedalus*, 112, 29–48.
- National Assessment Governing Board (NAGB). (2010). Science framework for the 2011 National Assessment of Educational Progress. ED 512544. Washington, DC: U.S. Department of Education.
- National Research Council. (1996). National Science Education Standards. Washington, DC: National Academy of Science Press.
- National Research Council. (2002). Scientific research in education (Committee on Scientific Principles for Education Research. In R. J. Shavelson & L. Towne (Eds.), Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.
- National Research Council. (2012). A framework for K-12 science education: Practices, crosscutting concepts, and core ideas. Committee on a conceptual framework for new K-12 science education standards. Board on Science Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.
- National Science Teachers Association. (1991). Position statement. Washington, DC: Author.
- Norris, S. P., & Phillips, L. M. (2003). How literacy in its fundamental sense is central to scientific literacy. *Science Education*, 87(2), 224–240.
- Organisation for Economic Co-operation and Development. (2003). The PISA 2003 assessment framework—mathematics, reading, science and problem solving: Knowledge and skills. Paris: Author.
- Organisation for Economic Co-operation and Development. (2006). Assessing scientific, reading, and mathematical literacy. Paris: Author.
- Organisation for Economic Co-operation and Development. (2007). PISA 2006: Science competencies for tomorrow's world, volume I analysis. Paris: Author.
- Osborne, J., Simon, S., & Collins, S. (2003). Attitudes towards science: A review of the literature and its implications. *International Journal of Science Education*, 25(9), 1049–1079.
- Pajares, F., & Valiante, G. (1997). Influence of self-efficacy on elementary students' writing. *Journal of Educational Research*, 90, 353–360.
- Perkins, D. N., Jay, E., & Tishman, S. (1993). Beyond abilities: A dispositional theory of thinking. *Merrill-Palmer Quarterly*, 39, 1–21.
- Popham, W. J. (1997). Consequential validity: Right concern—wrong concept. *Educational Measurement: Issues and Practice*, 16, 9–13.
- Presser, S., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., Rothgeb, J. M., et al. (2004). Methods for testing and evaluating survey questions. *The Public Opinion Quarterly*, 68, 109–130.
- Reynolds, C. R., Livingston, R. B., & Wilson, V. (2006). Measurement and assessment in education. Boston: Pearson.
- Roberts, D. A. (2007). Scientific literacy/science literacy. In S.K. Abell & N.G. Lederman (Eds.), *Handbook of research on science education* (pp. 729–780). Mahwah, NJ: Erlbaum.
- Ryder, J. (2001). Identifying science understanding for functional scientific literacy. *Studies in Science Education*, 36, 1–44.
- Shamos, M. (1995). The myth of scientific literacy. New Brunswick, NJ: Rutgers University Press.
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52, 591–611.
- Shaw, D., & Young, S. (2004). Revised guidelines for conducting item analyses of classroom tests. *The Researcher*, 18, 15–22.
- Shell, D. F., Colvin, C., & Bruning, R. H. (1995). Self-efficacy, attribution, and outcome expectancy mechanisms in reading and writing achievement: Grade-level and achievement-level differences. *Journal of Educational Psychology*, 87, 386–398.
- Shen, B. S. P. (1975). Science literacy and the public understanding of science. In S. B. Day (Ed.), *Communication of scientific information* (pp. 44–52). Basel, Switzerland: S. Karger A.G.
- Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16, 5–8, 13, 24.
- Showalter, V. M. (1974). What is united science education? Part 5. Program objectives and scientific literacy. *Prism II*, 2, 3–4.

- Sireci, S. G. (2013). Agreeing on validity arguments. *Journal of Educational Measurement*, 50, 99–104.
- Songer, N. B., & Linn, M. C. (1991). How do students' views of science influence knowledge integration? *Journal of Research in Science Teaching* 28: 761–784.
- United Nations Educational, Scientific and Cultural Organization. (1993). Final report: International forum on scientific and technological literacy for all. Paris: Author.
- Wenning, C. J. (2006). Assessing nature-of-science literacy as one component of scientific literacy. *Journal of Physics Teacher Education Online*, 3(4), 3–10.
- Wenning, C. J. (2007). Assessing inquiry skills as a component of scientific literacy. *Journal of Physics Teacher Education Online*, 4, 21–24.
- Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, 25, 68–81.
- Yore, L. D., Pimm, D., & Tuan, H. (2007). The literacy component of mathematical and scientific literacy. *International Journal of Science and Mathematics Education*, 5(4), 559–589.
- Zimmerman, C., Bisanz, G. L., Bisanz, J., Klein, J. S., & Klein, P. (2001). Science at the supermarket: a comparison of what appears in the popular press, experts' advice to readers, and what students want to know. *Public Understanding of Science*, 10, 37–58.

Science
Education

Using Rasch Measurement for the Development and Use of Affective Assessments in Science Education Research

TONI A. SONDERGELD,¹ CARLA C. JOHNSON²

¹*School of Educational Foundations, Leadership and Policy, College of Education and Human Development, and Center of Assessment and Evaluation Services (CAES), Bowling Green State University, Bowling Green, OH 43403, USA;* ²*Department of Curriculum and Instruction, College of Education, and Center for Advancing the Teaching and Learning of STEM (CATALYST), Purdue University, West Lafayette, IN 47909, USA*

Received 19 June 2012; accepted 7 March 2014

DOI 10.1002/sce.21118

Published online 5 June 2014 in Wiley Online Library (wileyonlinelibrary.com).

ABSTRACT: With the demand for quality quantitative instruments in the field of science education rising, additional measures of currently unassessed affective variables need to be constructed. In this study, we discuss the survey creation and evaluation process of the STEM Awareness Community Survey (SACS) through an application of Liu's (2010) framework for developing and using new affective instruments in science education. Liu's (2010) survey development framework uses Rasch measurement methods in survey evaluation to ensure a psychometrically sound measure is created. In surveying K-12 teachers, community business members, and higher education faculty through two iterative pilot studies, a unidimensional construct of STEM awareness and support was empirically established after misfitting items were removed and scale modifications were made. Further instructions on how to use results produced from the SACS are provided, so that even those with no or limited knowledge of Rasch analysis could use the instrument and interpret their findings with respect to the construct of STEM awareness and support. © 2014 Wiley Periodicals, Inc. *Sci Ed* **98**:581–613, 2014

Correspondence to: Toni A. Sondergeld; e-mail: tsonder@bgsu.edu

INTRODUCTION

A strong case for the importance of and need for developing quality science education measurements has been made by Liu (2010) in his book entitled *Using and Developing Measurement Instruments in Science Education: A Rasch Modeling Approach*. His argument is based on a documented research shift in the field to the use of more quantitative methods. According to Liu, this shift to more quantitatively focused science education research, and thus the demand for greater measurement development, is supported by the National Association for Research in Science Teaching's (NARST) membership communications in online forums, and findings from a 2009 NARST presidential sponsored survey of its membership which identified measurement as a top 10 grand challenge for science education research (Czerniak, 2009).

To meet the needs of science educational researchers asking quantitatively based questions, a greater number of measurement instruments including cognitive and affective assessments need to be developed and field tested. While classical test theory (CTT) methods have frequently been implemented in the instrument development literature in the field of science education (Liu, 2010), using Rasch (1980) methods is considered by many to be one of the best approaches for measure creation and refinement in any progressive field within the social sciences (Bond & Fox, 2007; Boone, Townsend, & Staver, 2010; Liu, 2010; Smith, 1996; Smith, Conrad, Chang, & Piazza, 2002; Waugh & Chapman, 2005; Wright, 1996). However, many educational researchers in general (including science educational researchers) lack the training needed to use Rasch measurement to inform instrument development (Liu, 2010; Smith et al., 2002).

While educational researchers may use Rasch methods less frequently than CTT, there has been a relatively recent push to introduce researchers to various Rasch measurement methods. Specifically in science educational research, *Science Education* has led the way by publishing three such introductory articles within the past decade, all of which have focused on different applications of Rasch measurement. First, Boone and Scantlebury (2006) showed how to apply the Rasch model for dichotomous data (Rasch, 1960) when examining a state-level multiple-choice test in science. The Rasch partial credit model (Masters, 1982) was then employed by Eggert and Bögenholz (2009) for developing an assessment of students' socioscientific decision-making strategies. And finally, in 2010 Boone, Townsend, and Starver illustrated how to assess a previously made survey of self-efficacy (STEBI [Science Teaching Efficacy Belief Instrument]; Enochs & Riggs, 1990) using the Rasch rating scale model (Andrich, 1978). Boone et al. (2010) further explained how the same principles for assessing a previously developed survey could also be applied to developing a new survey using the Rasch rating scale model.

The purpose of our paper is to add to these three introductory Rasch measurement pieces by demonstrating how to design a new instrument for assessing attitudes and beliefs using the Rasch rating scale model. In doing this, science educational researchers will be able to use these papers collectively as a primer for learning to address various measurement issues in their own work, and before tackling more thorough Rasch textbooks (e.g., Bond & Fox, 2007). As such, the STEM Awareness Community Survey (SACS) was created using Liu's (2010) framework for developing instruments to measure affective variables in science education. Specific research questions we are examining are as follows:

1. What empirically driven modifications should be made to the original SACS?
2. How do these changes affect the instrument's psychometric performance?
3. How is the construct of STEM awareness and support measured according to the SACS?

SURVEY RESEARCH AND DEVELOPMENT

Survey research is the most widely accepted method implemented for collecting data on individual attitudes, perceptions, and beliefs in the social sciences. Furthermore, Likert scales are the most commonly used scales in survey research when the goal is to measure a construct or variable. CTT techniques, such as a variety of different factor analysis methods, are often used in the social sciences for measuring and evaluating psychological constructs and personality assessments. However, Rasch measurement methods have been demonstrated to be advantageous over CTT in variable construction and validity assessment (i.e., Smith, 1996; Smith et al., 2002; Waugh & Chapman, 2005; Wright, 1996).

Although Rasch methods' value over CTT has been empirically evidenced, CTT continues to dominate the social sciences literature when it comes to survey development and analysis. It has been suggested by Smith et al. (2002) that this failure to use more advanced measurement methods (i.e., Rasch measurement) instead of CTT in psychological and personality scale construction and assessment is due to two main reasons. First, CTT assumptions are considered "weak," meaning they can be easily met with most data and the method is thus applicable in a wide variety of situations making CTT methods attractive to researchers. Second, Rasch measurement theory is typically only taught in advanced measurement courses to few graduate students, where CTT is taught in introductory quantitative research classes to nearly all. Science education researchers are no exception as very few doctoral programs in this field offer courses specifically focusing on measurement methods (Liu, 2010). This lack of measurement training in the social sciences may in part be due to the fact that CTT methods have existed since the early 1900s, and Rasch measurement methods are merely four decades old. Furthermore, science education researchers only began to use Rasch measurement methods for instrument development beginning in the early 1990s (Liu, 2010). Regardless of the reason why Rasch methods are less frequently implemented, a brief overview of the most important aspects of CTT and Rasch measurement with their major assumptions/specifications is provided below. For a more thorough introduction to Rasch measurement, see Bond and Fox (2007).

Classical Test Theory

With CTT, it is assumed that a measure of a person's attitudes or beliefs on a rating scale assessment form a linear combination of the total items. Each item on the survey is thus assumed to have the same difficulty and the same standard error. Therefore, it is common when analyzing rating scale data with CTT methods to sum scores on all items to compute a total score, where these scores are assumed to be at an interval level.

Based on the assumptions and capabilities of CTT measurement of rating scale data, many limitations to this method have been noted (e.g., Bode & Wright, 1999; Reise & Henson, 2003; Smith et al., 2002; Waugh & Chapman, 2005). Our brief discussion will not address all CTT limitations but will discuss those which are considered most important. First, rating scale data (e.g., Likert scales) are at the ordinal level. These scales are ranked, whereas the rank increases so does the trait of interest. However, it cannot be assumed that the increase in rank implies an equivalent increase in trait across rankings with ordinal level data. What this means with a typical 5-point Likert scale is that the difference between agree and strongly agree may not be the same as the difference between disagree and strongly disagree or disagree and neutral. Therefore, "raw score differences between pairs of points do not necessarily imply equal amounts of the construct under investigation" (Smith et al., 2002, p. 190). Only interval and ratio level data may be used for true measurement or with parametric tests (i.e., regression, analysis of variance (ANOVA), etc.). However, with CTT

analysis of rating scales, the data are often assumed to be interval level and possibly instead misused with parametric statistical procedures (Bode & Wright, 1999) if this assumption is not empirically verified.

As previously mentioned, each item on a rating scale assessment is assumed to possess the same amount of the trait being measured in CTT. This implies that regardless of how challenging an item is to endorse, it carries the same amount of weight toward a person's overall score on a rating scale assessment. For example, a survey participant might respond *strongly agree* to the items "Science is an important subject for primary school students" and also to the item "All primary school students need explicit science instruction daily to be successful in their future." With CTT, the weighting for these *strongly agree* responses would be identical although the latter item is clearly more challenging for people to endorse in comparison to the former item. CTT methods, however, do not take this notion of item difficulty into consideration when computing person measures. Additionally, if any item is removed or added to the assessment, the survey will result in a scale that is psychometrically different from the original. This item dependency on surveys in CTT prevents scores from the same set of items from being compared if any items are left incomplete unless some form of missing data imputation is performed. It is also difficult to compare performance from different assessment forms possessing different items or number of items.

Furthermore, CTT statistics such as reliability, standard error, means, and standard deviations are all dependent on the sample taking the survey. This means that an assessment could produce reliable results for one group completing the survey, but not for another being evaluated on the same construct (Hambleton, 2000). Factor structures from principal components or factor analysis may appear different depending upon the sample, making it difficult to compare results from various studies due to sample dependency.

Rasch Measurement Method

"Rasch models are mathematical models that require unidimensionality and result in additivity" (Smith et al., 2002, p. 190). To meet the unidimensionality specification for all Rasch models, a set of items must measure only a single construct or latent trait. In Rasch measurement, data must fit the mathematical model. And, if they do not, new data must be obtained. With regard to unidimensionality, if items are not measuring the same latent trait (as indicated by Rasch fit statistics) they need to either be eliminated or modified to better fit the model. While this specification of unidimensionality is a theoretical underpinning of measurement theory in general, it is very strictly adhered to with Rasch methods, which is oftentimes considered a criticism of Rasch measurement as data that do not fit the model must be abandoned and theory reconsidered.

When data do fit the model, the raw scores are converted into measurement units (logits—logarithm of odds) that are of interval level. Rasch measures allow both person ability and item difficulty for a latent trait to be placed on a single continuum so that people and items can be "estimated together in such a way that they are freed from the distributional properties of the incidental parameter, if the data fit the model" (Waugh & Chapman, 2005, p. 81). Unlike CTT, because of the interval level data and conjoint measurement scale between person ability and item difficulty, Rasch measurement indices are considered item and sample independent. Regardless of the sample chosen or items selected for assessment, as long as they are measuring the same construct, results may be obtained that are comparable across samples and various assessment forms within standard error estimates or confidence bands (Bond & Fox, 2007).

Missing data are no longer problematic as with CTT because of the probabilistic nature of Rasch measurement. This means that a properly constructed measurement can differentiate

between a person who endorses only the four most difficult items on survey and leaves the remaining blank and a person who endorses only the four easiest items on a survey while leaving the remaining blank. In CTT, these people would be given the same raw score of four and treated as equivalent since each item is assumed to be weighted the same. However, with Rasch measurement, the person endorsing the more difficult items would be rated higher than the person endorsing only the easier items as their missing data points would be probabilistically estimated with relation to the item difficulty of their endorsed responses.

DEVELOPING AFFECTIVE MEASURES IN SCIENCE EDUCATION USING RASCH

As stated earlier, Liu's (2010) framework for *Developing Instruments for Measuring Affective Variables* in science education research is used to structure the survey development process of the SACS instrument for assessing teachers, higher education faculty, and business community members' STEM awareness and support levels. The framework we have chosen to follow is similar to one proposed by Boone et al. (2010). And, both frameworks are strongly built upon two seminal works in the field of Rasch measurement that have been the impetus for pushing the social sciences in the direction of developing quality instruments before conducting statistical analysis: *Best Test Design* (Wright & Stone, 1979) and *Rating Scale Analysis* (Wright & Masters, 1982). Liu's survey development framework can be summarized as six major components: (1) identify purpose(s), (2) define construct, (3) initial survey development, (4) field testing, (5) implement Rasch measurement, and (6) develop guidelines for use. Some of the summarized framework components are composed of multiple pieces, all of which are first described in general below. After broadly explaining the components of Liu's framework, we demonstrate how to specifically apply this framework in developing a new affective instrument for science educational research.

Identifying the purpose(s) of the instrument provides justification for the development of a new affective measurement. This first component requires the researcher to look at literature and instruments previously published to see whether a tool currently exists to meet the researcher's needs or whether a new measurement is in fact needed. Additionally, the researcher identifies why the survey is being created and how the results should be used.

Once the purpose of the instrument has been determined, the researcher needs to focus on clearly defining the construct of interest. When defining a construct, the researcher needs to rely on theory as a guide for instrument development. Additionally, when identifying behaviors or attitudes that will be assessed as a proxy for the construct since such abstractions cannot be directly observed, the researcher must keep in mind that any construct or variable is required to assess varying levels of ability. Thus, a wide range of item difficulties (easier to endorse to more challenging to endorse) need to be created to measure the full range of the affective variable. While constructs can cover a wide range of attributes when measuring affective variables, it is critical to have all behaviors being assessed work toward clearly defining a single unidimensional measure. For example, a survey of high school students' chemistry self-efficacy may ask questions about a student's beliefs about their perceived abilities related to chemical reactions, ionic and covalent bonding, elements, balancing equations, atomic structure, and possibly algebra since mastering these content areas are all directly related to being successful in a typical chemistry class. However, the instrument probably would not ask questions about a student's perceived abilities related to ecology as these skills may not be thought of as directly related to a high school student's

chemistry ability, and thus may not fit well with the overall student's chemistry self-efficacy measure making it multidimensional.

The next component in the affective science education instrument development process uses multiple steps to generate initial survey items. From the theoretical elements that make up the construct, a table of specifications is developed to determine the number and type of items an instrument will have per behavior being assessed. An initial pool of items is then created based on these instrument specifications. Newly created items should then undergo expert review by content and methodological experts with the purpose of assessing item clarity and alignment with the measure's purpose and theoretical construct. Science education researchers may also choose to have typical participants the instrument is designed for complete cognitive interviews to help modify items for clarity of understanding.

After items are developed and have undergone expert review and/or typical participant cognitive interviews, the instrument should be field tested with a small sample representative of the intended population of interest. While larger sample sizes lead to more stable parameter estimates when analyzing the field test data with Rasch measurement techniques, a minimum goal is to have at least 10 participants per scale category. Therefore, a sample size as small as 50 participants is adequate for field testing when developing a traditional 5-point Likert-type scale.

From the data collected in field testing, Rasch analysis of the affective measure is conducted next. Various fit indices should be examined for items and persons to assess data to model fit and unidimensionality of the construct. Scale options (categories) need to be evaluated for their appropriateness in number. Validity and reliability of the overall construct should also be studied through the use of Rasch-generated figures and indices. Greater detail about specific Rasch statistics and fit indices are provided in the Methods section of this paper. Furthermore, Step 4 (field testing) and Step 5 (implement Rasch measurement) are often conducted multiple times in an iterative process to produce the highest quality measure.

Finally, it is necessary for the researcher to create and share guidelines for using the affective instrument that has been developed and evaluated. This last component of the affective instrument development process includes documentation of all previous components along with score reporting and interpretation information. Since most researchers interested in using a science education research survey are not familiar with Rasch analysis, a raw score to Rasch measure (logits) conversion table is helpful. This table allows researchers who are inexperienced with Rasch measurement to easily use the affective instrument and compare results from individuals in their sample to the construct that has been established.

APPLICATION OF LIU'S AFFECTIVE INSTRUMENTATION DEVELOPMENT FRAMEWORK

From this point forward, we demonstrate how to specifically apply the science education research affective instrument development process established by Liu (2010). We do this by sharing the development and evaluation process of the SACS created to assess the measure of STEM awareness and support by teachers, higher education faculty, and community business partners.

Component 1: Identifying Purpose(s) of the Instrument

Federal, state, and local investments into STEM (science, technology, engineering, and mathematics) education policy have increased significantly in recent years including a

2011 federal budget with \$4.3 billion appropriated for Race to the Top competitive grant funding, a program which positions STEM as the only competitive preference priority. Without a doubt, STEM has become a priority for the nation evidenced by comments from President Barack Obama's *Preparing Our Children for the Future, STEM Education in the 2011 Budget Report* where the President expressed his commitment "to moving our country from the middle to the top of the pack in science and math education over the next decade" (February 1, 2010). Furthermore, STEM educational investments have increased based primarily upon data that projects careers in mathematics and science will include over 80% of the most rapidly growing occupations—many of which have been routinely filled by talent from abroad (Bureau of Labor Statistics, 2010).

Despite monetary and educational investments in STEM being at record high levels, little attention has been devoted to generating a common understanding of STEM (Breiner, Harkness, Johnson, & Koehler, 2012). In addition, working with business, K-12 schools, and/or institutions of higher education to establish a grassroots effort to help community members understand the importance of STEM regarding the future prosperity of the United States in general, and specifically the preparedness of children for careers of now and the future, has been nonexistent (Johnson, 2012). Along with the lack of attention to building awareness, there has been no effort to gauge current understandings, awareness, and stakeholder engagement in STEM (Johnson, 2012). Recent research has suggested community engagement is essential for implementing and sustaining reform programs (Zmuda, Kuklis, & Kline, 2004). Furthermore, if the U.S. educational system is to make a dramatic shift in how children are prepared for careers of tomorrow it will require collaborative efforts with multiple stakeholders to accomplish (Shirley, 2009). Therefore, it is critical to assess STEM awareness at multiple levels of the educational community and take necessary steps to address potential gaps in understanding. While the importance of measuring teachers, higher education faculty, and/or community business members' STEM awareness has been established, no current tool to do this exists, thereby justifying the creation of a new affective instrument.

METHODS

Component 2: Define Construct

The state this study was conducted in was awarded Race to the Top funding. An integral part of the application was the establishment of a network infrastructure to support reform of STEM education. A portfolio of STEM investments is included in the network strategic plan. The network devoted funding for an external evaluation of the STEM investments. Again, as STEM awareness has not been a focus of research to this point, there is a dearth of existing assessments aligned with this construct. Therefore, a new instrument was created for this research using the network strategic plan goals as a guiding framework for development. The network strategic plan goals include (1) increase student interest, participation, and achievement in STEM; (2) expand student access to effective STEM teachers and leaders; (3) ensure a well-prepared, ready-made STEM workforce for the state by reducing the STEM talent and skills gap; (4) build community awareness and support for STEM. The instrument developed to assess STEM awareness and support aligns primarily on Goal 4, but also gleans some insight from other goals as well. This emphasis is purposeful, as this survey is the main data-gathering tool aligned with Goal 4, whereas other data sources inform progress on the other goals.

To define the construct of STEM awareness and support, 12 initial open-ended items aligned with network strategic plan goals were developed and sent out to participants in K-12

TABLE 1

Table of Specifications Indicating the Number of Items Needed From Each Thematic Component Making Up the Construct of STEM Awareness and Support

Thematic Component of Stem Awareness and Support	Number of Items Needed
Industry Engagement in STEM Education (IE)	12–15
STEM Awareness and Resources (AR)	12–15
Regional STEM Careers and Workforce (CW)	12–15
Preparation of Students for Success in College and Careers (PR)	5–10

education, business, and informal education agencies in two states (not including the actual state for this study). Based upon responses from the open-ended questionnaire, four main themes were identified: industry engagement in STEM education (IE), STEM awareness and resources (AR), regional STEM careers and workforce (CW), and preparation of students for success in college and careers (PR). These themes, or behaviors of the construct that need to be measured, were then used to drive the development of Likert-scale items for the new STEM awareness and support affective measure. In planning for item creation, a table of specifications was generated to align with thematic importance as indicated from the open-ended survey. Since IE, AR, and CW were all mentioned with similar frequency and PR was brought up less often, our table of specifications reflects this difference (see Table 1).

Additionally, open-ended survey results along with the goals of network informed the theory behind the construct of STEM awareness and support. It was hypothesized that items from each of the four themes (IE, AR, CW, and PR) would run along the full continuum of items with more general items being easier for participants to agree with compared to more specific items. For instance, considering AR, it was hypothesized that an item asking about the general *need for more work in spreading awareness of STEM education* would be easier for participants to agree with in comparison with a more specific item about *parents understanding the importance of STEM education*. Furthermore, it was expected that although items from each thematic group would run the full continuum of easy to difficult to agree with, IE and AR would have a greater number of easier items and CW and PR would have a greater number of more difficult items due to the nature of these themes. Figure 1 illustrates the theoretical model behind the construct of STEM awareness and support.

Component 3: Initial Survey Development

Three parallel versions of the SACS were created to assess K-12 teachers, higher education faculty, and members from the business community in their attitudes and beliefs about regional STEM awareness and support. The surveys were each composed of 63 items with the majority of items being on a traditional 1–5 point Likert-scale (*strongly disagree* to *strongly agree*), but also having some select all appropriate items, and open-ended questions. Seven sections were developed for the survey: demographic information (six items), employment/employer information (four items), industry engagement in STEM education (12 Likert-scale items), STEM awareness and resources (14 Likert-scale items), preparation of students for success in college and careers (six Likert-scale items and two select appropriate options item), regional STEM careers and workforce (13 Likert-scale items,

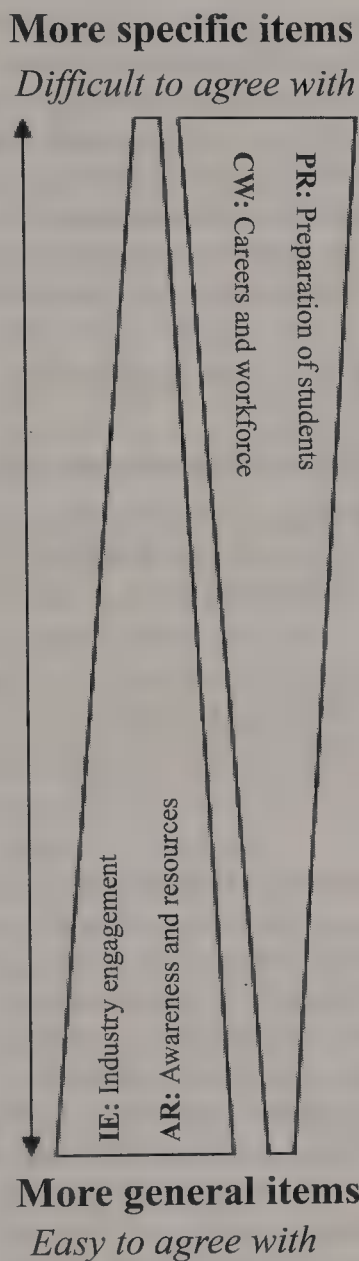


Figure 1. Theoretical model of the construct of STEM awareness and support. IE, AR, CW, and PR items are hypothesized to all run along the continuum of items with more general items being easier to agree with and more specific items more difficult to agree with. Many CW and PR items are more specific and thus anticipated to be more difficult to agree with. Many IE and AR items are more general and thus anticipated to be easier to agree with.

one select appropriate options item), and an open-ended questions section (five questions). Only four of the seven survey sections—those based on the identified theoretical themes—are used to measure the construct of STEM awareness and support: IE, AR, PR, and CW. Additionally, only the Likert-scale items from these sections are used in the analysis of the construct.

Component 4: Field Testing— Pilot 1 Survey Distribution

Sample and Procedures. All surveys were initially piloted with their appropriate audiences, but with individuals in another state outside of the state involved in this study, where

the SACS has been developed for use in evaluating Race to the Top STEM Education Programs. A convenience sample of 72 participants completed the initial pilot survey: 39 K-12 teachers, 17 higher education faculty, and 16 business community members. For field testing purposes, a sample size of 72 is appropriate for this instrument since a 5-point Likert-scale was used and the goal is to have a minimum of 10 participants per scale category, thus making 50 the minimum number of participants acceptable for this situation (Liu, 2010). The SACS survey items were put into Survey Monkey and emailed to the teachers who were participating in a master's level course at a university in Ohio. Higher education and business community members in the Ohio area were recruited through existing listservs and previous activity in regional educational programming.

Component 5: Implement Rasch Measurement—Pilot 1

The Rasch rating scale model (Andrich, 1978) for polychotomous responses is used in this survey evaluation study to investigate the overall fit of the data to the Rasch model, underlying unidimensionality of STEM awareness and support according to the SACS, individual misfitting items and people, and overall scale functioning. WINSTEPS version 3.74.0 (Linacre, 2012) was used for all Rasch analyses. Assessing the data fit to the Rasch model is a holistic and iterative process. Once unidimensionality, item and person fit, and scale functioning are assessed, understanding the construct according to the survey is then possible.

Rasch Item and Person Fit Statistics. Numerous Rasch fit indices are used to determine item and person fit in this study. Both infit and outfit statistics are assessed for items and persons. With regard to person attitudes and beliefs, infit and outfit indices provided information about unexpected patterns of responses. Outfit indicates irregular response patterns far from a person's attitude measure, and infit shows the existence of unexpected responses close to the person's attitude level.

Reasonable item/person mean-square infit and outfit ranges from 0.6 to 1.4 logits for rating scale or survey data (Wright & Linacre, 1994). Items/persons with a mean-square greater than 1.4 misfit and are providing more misinformation (noise) than information. While items/persons with a mean-square less than 0.6 overfit and their response patterns are considered too predictable and are therefore not adding any new information to understanding the construct. Overfitting items/persons are of less concern than those that misfit, since overfitting items/persons typically do not harm the measure.

Direction of point-biserial correlation (pbis) for each item with the overall construct is also assessed for item fit. Items with a positive point biserial are contributing to the measure, whereas items with a negative point biserial are acting in opposition to the measure's construct and thus must be removed as they do not fit with the variable. While infit, outfit, and point biserial are all taken into consideration when deciding if an item or person "fits" with the measure, there are no concrete criteria for determining whether a person or item should be removed from a survey completely. For purposes of this analysis, if an item or person has a negative point biserial it will be removed, and if a person or item exceeds the infit and outfit criteria described above they will be considered for removal if their exclusion improves the overall scale functioning as a result. Item/person fit assessment and removal is an iterative process and thus is done in multiple phases of analysis.

Person and Item Reliability and Separation. Separation and reliability indices for people and items are also used to determine the quality of measures. Rasch separation indicates the

number of statistically distinct groups that can be classified along a variable. Computing separation is essential because a measure “is useful only if persons differ in the extent to which they possess the trait measured” (Bode & Wright, 1999, p. 295). Rasch reliability is similar to traditional reliability in that it is the statistical reproducibility of a set of values. While traditional reliability is computed for raw scores, Rasch reliability is computed for person abilities and item difficulties. Separation and reliability of 1.50 and 0.70, respectively, are considered acceptable; 2.00 and 0.80, respectively, are good; 3.00 and 0.90, respectively, represent excellent levels (Duncan, Bode, Lai, & Perera, 2003).

Rating Scale Analysis. Further investigation of rating scale category counts, average and expected rating scale category measures, outfit mean-square statistics, and step calibrations are used to examine and optimize the functioning of the rating scale categories in the SACS. Linacre’s (2002) guidelines for optimizing rating scale categories are used as a general framework to assist with the empirical evaluation of the SACS. In line with Linacre’s (2002) guidelines for evaluating rating scale category effectiveness, the following criteria will be assessed of the SACS:

1. *At least 10 observations in each rating scale category.* Step calibration is imprecisely estimated and potentially unstable when observed frequency is low in a category.
2. *Average measures advance monotonically with category.* Higher categories need to be produced by higher measures otherwise we do not know the meaning of the rating scale, and it will render the measures useless.
3. *Outfit mean-squares less than 2.0.* Values greater than 2.0 indicate there may be more unexplained noise than explained noise. This suggests more misinformation is being obtained than information.
4. *Step calibrations advance by at least 1.4 logits for a 5-point scale or 1.0 for a four-point scale, but less than 5.0 logits regardless of scale categories.* If step calibrations are 1.4 logits or greater for a 5-point scale or 1.0 logits or greater for a 4-point scale, it implies the rating scale is composed of subtests equal to the number of categories. Therefore, if a person is in a high category, they have successfully met the requirements of the lower categories. However, if the step calibrations are 5.0 or higher, the category boundaries are too far apart resulting in a “dead zone.”

Multiple analysis iterations are required for assessing and developing the best survey items to measure a construct or variable. To do this, a holistic lens must be used assessing all components described above. Thus, the results section is divided into two main parts. The Analysis I section details the scale and item fit iterations conducted to determine the best functioning scale with only the items fitting the Rasch model to, if possible, obtain construct unidimensionality. Data from fitting items and people on an appropriate scale are then used in the Analysis II section to further investigate the meaning of the STEM awareness and support construct and assess the variable itself for redundancy or gaps in the measure. All of this information is then taken into consideration, the survey is revised, and a second distribution of the modified instrument is completed with all analyses conducted again.

Analysis I: Scale and Item/Person Fit Iterations—Pilot 1. A summary of Analysis I results are described in the text, whereas Table 2 provides a detailed comparison of each criteria being examined. Initially, the analysis was run with all persons and items on the

TABLE 2
Multiple Scale and Item Analysis Iterations Results From Pilot 1 and Pilot 2

Guideline/criteria	Pilot 1 Iterations		Pilot 2 Iterations	
	Original 1–5 scale (SD, D, N, A, SA) <i>All items and persons</i>	Collapsed 1–4 scale (SD, D/N, A, SA) <i>All items and persons</i>	Collapsed 1–4 Scale (SD, D/N, A, SA) <i>Removed six misfitting items (IE5, IE6, AR8, CW6, AR7, CW10)</i>	Revised 1–4 Scale (SD, D, A, SA) <i>All items and persons</i> <i>Revised 1–4 Scale (SD, D, A, SA) <i>Removed six misfitting persons</i></i>
Scale function: At least 10 observations of each category	Acceptable	Acceptable	Acceptable	Acceptable
Scale function: Average measures advance monotonically with category	Acceptable	Acceptable	Acceptable	Acceptable
Scale function: Outfit mean-squares less than 2.0	Acceptable	Acceptable	Acceptable	Acceptable
Scale function: Step difficulties advance by at least 1.0 logits for a 5-point scale; 1.4 for 4-point scale	Problematic between points 2 and 3 (only 0.73 difference)	Problematic between points 3 and 4 (only 1.19 difference)	Acceptable	Acceptable
Scale Function: step difficulties advance by less than 5.0 logits	Acceptable	Acceptable	Acceptable	Acceptable
Person reliability	0.87	0.88	0.87	0.92
Person separation	2.37	2.68	2.64	3.37
Persons—pbis	None	None	None	0
Item reliability	0.97	0.96	0.97	1.00
Item separation	5.36	5.16	5.64	15.89
Items—pbis	None	CW5	None	None
Items misfitting	IE5, IE6, AR7, AR8, CW9, CW10	IE5, IE6, AR7, AR8, CW6, CW10	None	None

Abbreviations: SD, *strongly disagree*; D, *disagree*; N, *neutral*, A, *agree*; SA, *strongly agree*; D/N, *disagree/neutral*.

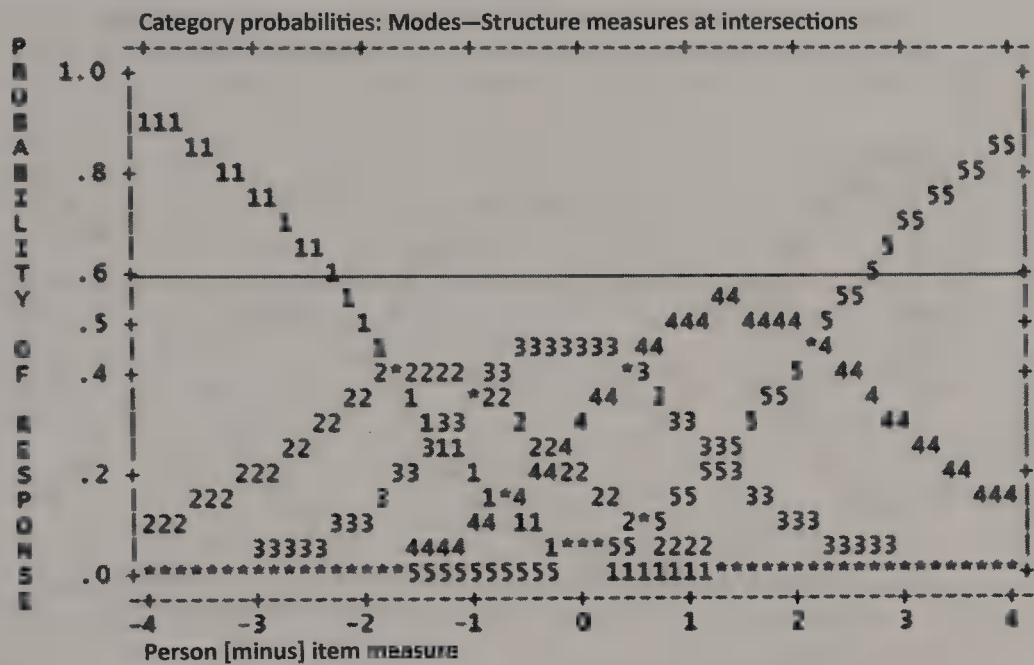


Figure 2. Scale functioning output from Winsteps for initial survey distribution run on 5-point scale (*strongly disagree, disagree, neutral, agree, strongly agree*). All probability curves should have their own peak which reaches approximately .6 probability of response for the scale to function well. Survey response options 2 (*disagree*) and 3 (*neutral*) are not meeting these criteria and should be collapsed or investigated further.

original 5-point scale (1 = *strongly disagree*, 2 = *disagree*, 3 = *neutral*, 4 = *agree*, 5 = *strongly agree*). All rating scale assessment components were acceptable with the exception of the step difficulties. The steps should advance by a minimum of 1.4 logits for a 5-point scale, and between the *disagree* (2) and *neutral* (3) categories the steps only advanced by 0.73 logits. Figure 2 illustrates this issue graphically. The scale functioning output figure from Winsteps should show all probability curves for each response option with their own peak reaching approximately .6 probability of response for the scale to function well. Supporting the results from the step difficulty advancement not meeting minimal criteria for disagree and neutral, survey response options 2 (*disagree*) and 3 (*neutral*) are not meeting these criteria and should be collapsed or investigated further. With this model, six items were questionable regarding their fit indices and no persons misfit. Person and item separation and reliability were considered good and excellent, respectively.

Based on these initial results, it was determined that the scale should be collapsed combining the *disagree* (2) responses with the *neutral* (3) responses. This is often the case with 5-point scales; respondents frequently select neutral when they really disagree, and these categories therefore overlap in their meaning. After running the analysis with all items and persons again, but collapsing the *disagree* (2) and *neutral* (3) categories together a similar picture was revealed with some improvement. There was a new issue with the rating scale categories between the *agree* (3) and *strongly agree* (4) categories, but these were very close to being acceptable. Again, six items misfit and no persons misfit.

With both previous analysis results directing the third iteration of analyses, the six misfitting items were removed and the analysis run again with the 1–4 point scale having collapsed the *disagree* (2) and *neutral* (3) categories. After these modifications, all scale criteria were acceptable, no misfitting items remained (indicating a unidimensional construct), item separation and reliability were the highest of any run, and person

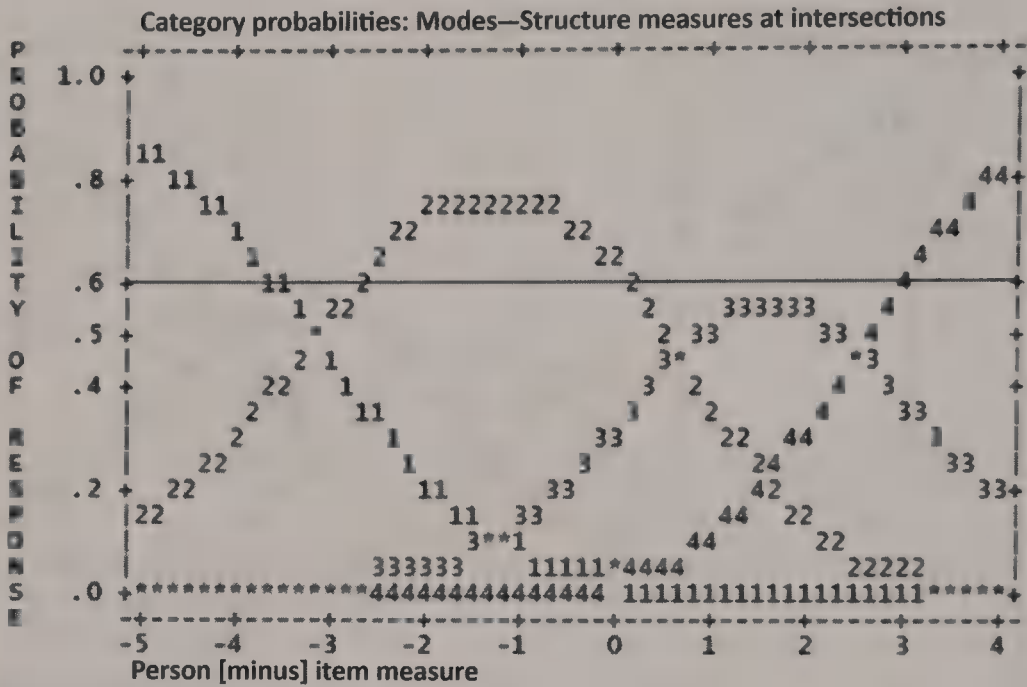


Figure 3. Scale functioning output from Winsteps for initial survey distribution run on 4-point scale (*strongly disagree, disagree/neutral, agree, strongly agree*). All response option probabilities have their own distinct peak, which reaches approximately .6 probability of response suggesting a well functioning scale.

separation and reliability were similar to the other runs. Figure 3 illustrates the scale was functioning well since all response option probabilities have their own distinct peak, which reaches approximately .6 probability of response. Therefore, it was determined that this item and scale arrangement was best for assessing the construct of STEM awareness and support.

Analysis II: STEM Awareness and Support Construct Meaning—Pilot 1. All results discussed in this section are in reference to Figure 4—the Rasch variable map of the SACS. A variable map, also called a Wright map named after Benjamin Wright (Wilson & Draney, 2000), graphically illustrates the conjoint ruler of the construct of interest by depicting the distribution of items on the right-hand side of the map and the distribution of people on the left-hand side. Items are arranged by difficulty from items that are easiest to agree with at the bottom of the ruler and more difficult items to agree with at the top of the ruler. People are arranged by attitude/belief level with those having the lowest attitudes/beliefs on the bottom of the ruler and those with the highest attitudes/beliefs toward the top of the ruler. The unit of measure on this ruler for both items and people is a logit, which is a log odds transformation of the probability of endorsing a survey item. Items that are at the same logit measure as a person have a 50% probability of being endorsed by that person. Items below the person’s trait measure have a greater probability of being agreed with, whereas items above a person’s trait measure have a lower probability of being endorsed.

Although people and items are aligned on the same ruler, they often do not have the same mean. The item mean difficulty is always fixed at 0 logits regardless of the construct being measured, and the person mean attitude/belief is established in comparison with the average item difficulty. If the item mean difficulty is above the person mean attitude/belief, then the survey in general is considered difficult to agree with for the group of people taking

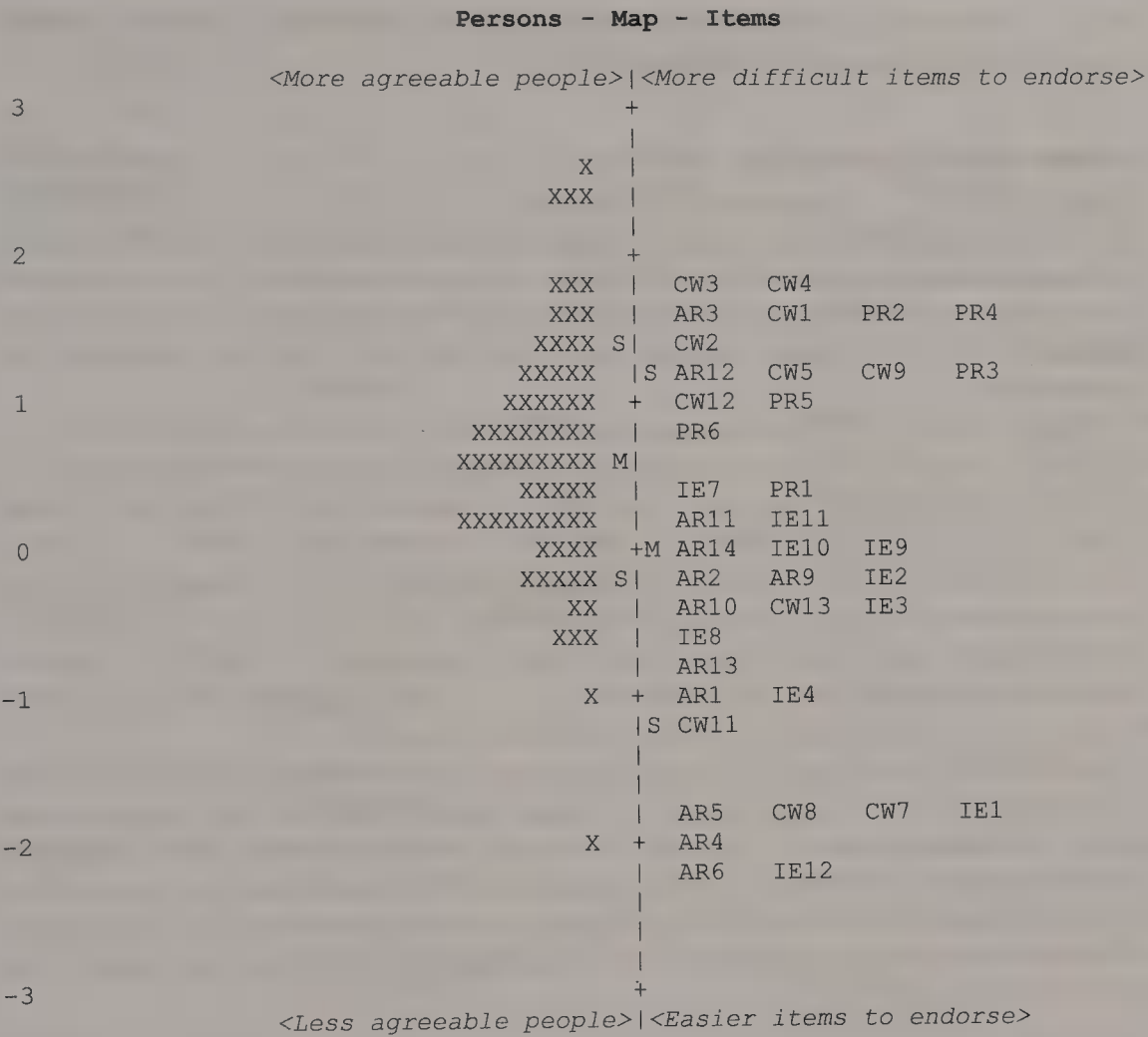


Figure 4. Pilot 1 variable map of the construct of STEM awareness and support as measured by the SACS. Individual respondents are shown on the left-hand side of the map in comparison to the items on the right-hand side of the map. Each “X” represents one respondent.

the survey. Conversely, if the mean item difficulty is below the person mean attitude/belief the survey is considered easy for the average person taking the test to agree with. Having a survey with the item and person mean at the same place (0 logits) on the ruler is ideal. When looking at Figure 4, it is evident that the item difficulty mean (0 logits) is less than the person attitude mean (0.63 logits) by nearly 2/3 of a logit. This indicates the SACS items are relatively closely aligned with the respondents taking this survey although the items are slightly easier than the average persons’ attitudes toward the construct.

A variable map is a unique tool that allows researchers to assess the validity of their instrument in respect to the definition of the construct. Important validity related issues to examine by the variable map are item coverage of the variable’s full range, theoretically based definition of the construct according to the variable map, and linear progression along the continuum of items (Liu, 2010). The SACS was created with the intent of assessing four components in the construct of STEM awareness and support: IE, AR, PR, and CW. As described earlier, items for each component of the construct were developed to measure easier (general) and more difficult (specific) to agree with beliefs/attitudes toward STEM awareness and support. Additionally, we hypothesized that more IE and AR items would be easier for our participants to agree with since more were general in nature, and a greater number of PR and CW items would be more challenging to endorse as they were more

specific. This theoretical conceptualization of the STEM awareness and support construct was supported as shown in the variable map.

Possible Scale Improvements. As previously noted in the Analysis I section for the initial survey pilot, six items were removed due to misfit issues (IE5, IE6, AR8, CW10, CW6, AR7). In addition to removing misfitting items, it is also common to remove items that are redundant in both content and difficulty to maximize parsimony. And, it is common to add items where there are gaps in the continuum to ensure all parts of the variable are being well measured. To make such decisions the variable map in Figure 4 is again used as a guide.

Redundancy of item difficulty measuring STEM awareness and support around the -2 logit range is shown by multiple items measuring at the same difficulty within standard error. Item IE12 should be removed because the content is redundant with IE1, and these items are also statistically similar in difficulty. The choice to remove IE12 over IE1 was because IE1 refers directly to STEM community partnerships, while IE12 is broader referring to community partnerships in general. At the 0 logit range, items IE9 and IE10 are similar in difficulty and also assessing similar content, thus it is recommended that IE9 be removed because it is less specific in comparison with IE10. All items suggested for removal from the survey are listed in Table 3 with their appropriate rationale.

Items on this scale appear slightly easier to endorse than the participants' attitudes toward the construct. This is evidenced by the item measure mean at 0 logits and the person measure mean at 0.63 logits. While this slight difference exists, when taking the standard deviations of the item measures (1.02 logits) and person measures (0.67 logits) into consideration, we see that the means are not statistically different as they fall within ± 2 SD of each other. Thus, items are doing an acceptable job of targeting and measuring Pilot 1 participant STEM awareness and support beliefs.

Component 4: Field Testing—Pilot 2 Revised Survey Distribution

Sample and Procedures: Pilot 2. After instrument modifications were made from the initial survey pilot, the revised SACS was redistributed and data collected from 600 participants (200 from each group—K-12 teachers, higher education faculty, and business community members) were used for a second pilot. This time, the SACS was composed of 39 total Likert-scale items (rather than the initial 47) from the following sections: 8 IE, 13 AR, 6 PR, and 12 CW. All Likert-scale items were now on a 4-point scale consisting of *strongly disagree* (1), *disagree* (2), *agree* (3), and *strongly agree* (4) rather than a traditional 5-point scale. The SACS survey items were again put into Survey Monkey and emailed to participants. K-12 teachers and higher education faculty were selected from state databases, and business community participants were recruited from lists provided by regional chambers of commerce.

Component 5: Implement Rasch Measurement—Pilot 2

A summary of Analysis I results is given in the text, whereas Table 2 provides a detailed comparison of each criteria being examined. Initially, the analysis was run with all persons and items on the revised 4-point scale (1 = *strongly disagree*, 2 = *disagree*, 3 = *agree*, 4 = *strongly agree*). All rating scale assessment components were acceptable; person and item reliabilities and separations were all considered excellent. No items were misfitting, but six persons had negative point biserials.

TABLE 3
Misfitting Items from Initial Rasch Run of the SACS Analysis (Pilot 1)

Analysis Phase Removed and Item Code	Item Text	Problematic Issue(s)	Recommendation and Rationale
Analysis I: IE5	I have been involved in coteaching STEM lessons with community/business members.	Infit > 1.4 Outfit > 1.4	IE5 and IE6 are similar in that seem very “classroom” centered. Perhaps this is why they do not fit well with the construct because all other collaborative items are broader in scope. Remove these two items from this analysis and future survey administrations.
Analysis I: IE6	I have been involved in STEM curriculum planning with community/business stakeholders.	Infit > 1.4 Outfit > 1.4	
Analysis I: AR8	Students must complete a 4-year college degree to secure a career in a STEM field.	Infit > 1.4 Outfit > 1.4	Possibly the idea of a “4-year” degree is making this item misfit. Remove from this analysis but consider revising and trying on future survey administrations. Suggested revision: <i>Students with postsecondary education are more likely to secure a career in a STEM field than those without a college degree.</i>
Analysis I: CW10	STEM education is for all students.	Infit > 1.4 Outfit > 1.4	This item is very vague and could be interpreted differently by different respondents. Remove from this analysis and consider revising and trying on future survey administrations. Suggested revision: <i>All K-12 students should have access to STEM education.</i>
Analysis I: CW6	K-12 educators and administrators are aware of the workforce needs of area employers.	Infit > 1.4 Outfit > 1.4 -pbis	Remove from analysis and future surveys administrations because this item misfits on all three indices.

(Continued)

TABLE 3
Continued

Analysis Phase Removed and Item Code	Item Text	Problematic Issue(s)	Recommendation and Rationale
Analysis I: AR7	Students can be successful in college without a firm understanding of STEM subjects.	Infit > 1.4 Outfit > 1.4	Negatively worded items often misfit from constructs. Remove from this analysis and typically would revise to be positively phrased. However, AR5 is this item positively phrased so this item should be removed from all future analyses as well.
Analysis II: IE12	It is important for educators to build relationships with members of the outside community.	Redundant in content and difficulty	This item is similar in content and difficulty with item IE1: <i>I believe it is important for area businesses to be involved in STEM partnership(s) with K-12 schools in my region.</i> IE12 should be removed from the survey completely as it is not adding any additional information to the construct.
Analysis II: IE9	There has been an increase in K-12 STEM education opportunities offered by organizations within my region in the last year.	Redundant in content and difficulty	This item is similar in content and difficulty with item IE10: <i>Overall, there has been an increase in K-12 STEM education opportunities for students in the region in the last year.</i> IE9 should be removed from the survey completely as it is not adding any additional information to the construct and is less specific than IE10.

A second iteration of analysis was conducted removing the six misfitting persons. The removal of these persons produced results with all scale criteria acceptable, no misfitting items (indicating a unidimensional construct), item and person separation and reliability were the highest of any run (classified as excellent). Figure 5 illustrates the scale was functioning well since all response option probabilities have their own distinct peak which reaches at least .6 probability of response. Therefore, it was determined that this item and scale arrangement was best for assessing the construct of STEM awareness and support out of all pilot analyses.

Analysis II: STEM Awareness and Support Construct Meaning—Pilot 2. A variable map extremely similar to that of the Pilot 1 (Figure 4) was produced for Pilot 2 (Figure 6). Again, the theoretical structure of STEM awareness and support being measured by all

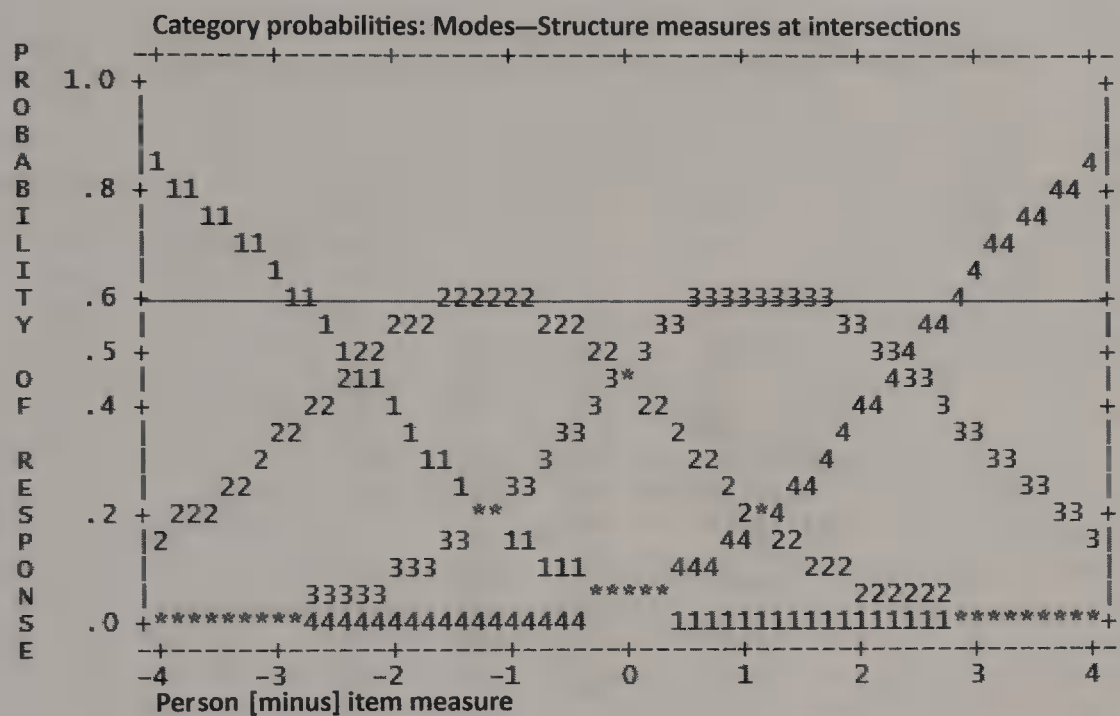


Figure 5. Scale functioning output from Winsteps for revised survey distribution run on 4-point scale (*strongly disagree, disagree, agree, strongly agree*). All response option probabilities have their own distinct peak which reaches a minimum .6 probability of response suggesting a well functioning scale.

four components—IE, AR, PR, and CW—along the continuum of the variable from easier (general items) to more difficult (specific items) to agree with beliefs/attitudes still existed. Furthermore, in comparing the 37 similar item measures from Pilot 1 to Pilot 2, the vast majority of items ($n = 35$; 94.6%) were in the same position. This was determined by comparing the logit measure of each item from each pilot. If the item measures were within ± 2 standard errors (SE) of each other then they were deemed statistically similar in terms of measurement position. Both items similar in content that were not in the statistically same measurement position were significantly more difficult for the Pilot 2 population to endorse compared to Pilot 1. (See Appendix A for the complete comparison of Pilot 1 to Pilot 2 item measures. Appendix B shows the final items on the SACS.)

Similar to Pilot 1, items on the Pilot 2 variable map appear slightly easier to endorse than the participants' attitudes toward the construct. Yet again, we see that when taking item and person measure means and standard deviations into consideration this difference is not statistically significant as the means overlap within $\pm 2 SD$ (item measure: $M = 0.00$ logits, $SD = 1.04$ logits; person measure: $M = 0.69$ logits, $SD = 0.91$ logits). This again supports the notion that the SACS items are targeting and measuring participant STEM awareness and support beliefs in an acceptable fashion.

DISCUSSION

Component 6: Developing Guidelines for Use

Using Rasch measurement methods revealed that the construct of STEM awareness and support is a unidimensional variable that fits the Rasch model after numerous iterations within two pilot tests. Knowing where a teacher, school, or community stakeholder falls on this ruler can provide practical value by informing where STEM reform programs should focus their efforts to scaffold participants to higher attitudinal levels. Additionally,

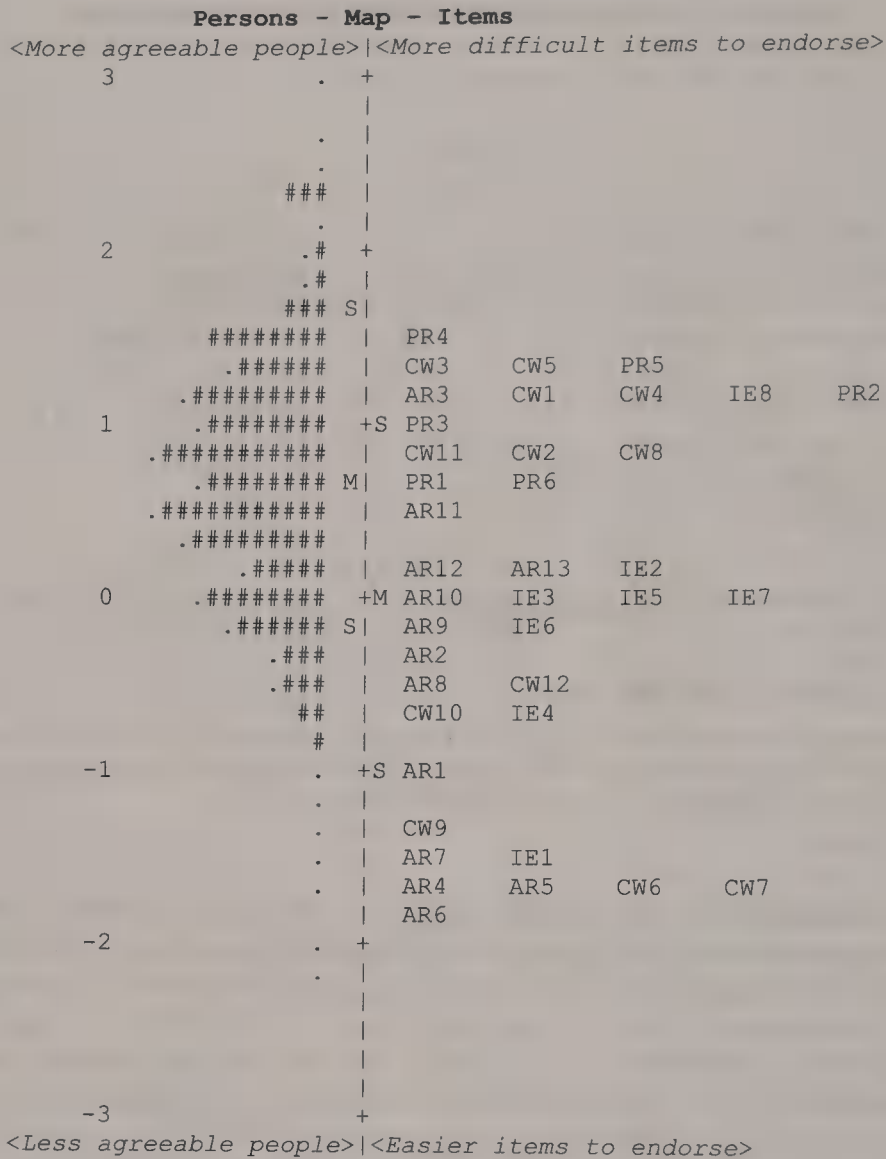


Figure 6. Pilot 2 variable map of the construct of STEM awareness and support as measured by the SACS. Respondents are shown on the left-hand side of the map in comparison to the items on the right-hand side of the map. Each “#” represents five respondents, and each “.” is between one and four respondents.

knowing where a community falls on this continuum may help in the building of stronger STEM partnerships between K-12 schools, higher education institutions, and local businesses by showing where specific needs exist and providing concrete goals to strive for to move up the ruler. Not only is it of great importance that there is a common understanding of STEM across these groups (Breiner et al., 2012), but more specifically community engagement is essential for implementing and sustaining reform programs (Zmuda et al., 2004). Moreover, it will take a village to prepare our children for the STEM careers of the future, and awareness is the first step to empowerment (Shirley, 2009).

Since a large number of science education researchers have not been trained in conducting Rasch analysis (Liu, 2010), but some may want to use the SACS and compare their participants to the construct’s measure, a special raw score to logits conversion is provided in Figure 7. To use this conversion table, a researcher would give the SACS revised final survey to participants, calculate their raw scores from all Likert-scale items (possible range

Score	Measure	SE	Score	Measure	SE	Score	Measure	SE
39	-7.69E	1.84	79	-1.18	0.26	119	1.37	0.26
40	-6.45	1.02	80	-1.11	0.25	120	1.44	0.26
41	-5.72	0.74	81	-1.05	0.25	121	1.51	0.27
42	-5.27	0.61	82	-0.99	0.25	122	1.58	0.27
43	-4.94	0.54	83	-0.92	0.25	123	1.65	0.27
44	-4.67	0.49	84	-0.86	0.25	124	1.72	0.27
45	-4.44	0.46	85	-0.80	0.25	125	1.80	0.27
46	-4.25	0.43	86	-0.73	0.25	126	1.87	0.27
47	-4.07	0.41	87	-0.67	0.25	127	1.95	0.28
48	-3.91	0.39	88	-0.61	0.25	128	2.02	0.28
49	-3.76	0.38	89	-0.54	0.25	129	2.10	0.28
50	-3.63	0.36	90	-0.48	0.25	130	2.18	0.28
51	-3.50	0.35	91	-0.42	0.25	131	2.26	0.29
52	-3.37	0.34	92	-0.36	0.25	132	2.34	0.29
53	-3.26	0.34	93	-0.29	0.25	133	2.43	0.29
54	-3.15	0.33	94	-0.23	0.25	134	2.51	0.29
55	-3.04	0.32	95	-0.17	0.25	135	2.60	0.30
56	-2.94	0.31	96	-0.11	0.25	136	2.69	0.30
57	-2.85	0.31	97	-0.05	0.25	137	2.78	0.31
58	-2.75	0.30	98	0.02	0.25	138	2.88	0.31
59	-2.66	0.30	99	0.08	0.25	139	2.98	0.32
60	-2.57	0.30	100	0.14	0.25	140	3.08	0.32
61	-2.49	0.29	101	0.20	0.25	141	3.18	0.33
62	-2.40	0.29	102	0.27	0.25	142	3.29	0.33
63	-2.32	0.28	103	0.33	0.25	143	3.41	0.34
64	-2.24	0.28	104	0.39	0.25	144	3.52	0.35
65	-2.16	0.28	105	0.45	0.25	145	3.65	0.36
66	-2.09	0.28	106	0.52	0.25	146	3.79	0.37
67	-2.01	0.27	107	0.58	0.25	147	3.93	0.39
68	-1.94	0.27	108	0.64	0.25	148	4.09	0.40
69	-1.86	0.27	109	0.71	0.25	149	4.26	0.43
70	-1.79	0.27	110	0.77	0.25	150	4.45	0.45
71	-1.72	0.27	111	0.84	0.25	151	4.67	0.49
72	-1.65	0.26	112	0.90	0.26	152	4.93	0.54
73	-1.58	0.26	113	0.97	0.26	153	5.25	0.61
74	-1.51	0.26	114	1.03	0.26	154	5.70	0.73
75	-1.44	0.26	115	1.10	0.26	155	6.42	1.02
76	-1.38	0.26	116	1.17	0.26	156	7.65E	1.84
77	-1.31	0.26	117	1.23	0.26			
78	-1.25	0.26	118	1.30	0.26			

Figure 7. Raw score to logit conversion table for 39 Likert-scale items on the final SACS instrument. To use this table, compute a participant's raw score from the survey (39–156) and match their score with the corresponding logit measure in the figure. Then find where this falls on the Rasch variable map in Figure 6 to compare the person's score with the construct.

between 39 and 156 exists for the 39 item instrument on a 1–4 point scale), and determine the appropriate corresponding logit score. Participant logit scores can then be plotted on the variable map provided in Figure 6 to make a comparison of their participants' attitude levels to the STEM attitudes and support construct. With CTT, this type of participant to previously developed scale comparison would not be appropriate because of sample dependency; however, it is made possible with the person and item free characteristics of Rasch measurement (Hambleton, 2000). Therefore, a key advantage of instruments developed using Rasch measurement methods is that they are able to be utilized by a much larger audience.

A practical example of how using the conversion table in collaboration with the final variable map could assist in moving a STEM reform effort forward is provided in the

following scenario: Perhaps a school district has been engaged in STEM reform efforts, and the main focus has been on improving teacher STEM content knowledge in the first year. Teachers complete the SACS at the end of the first year, and on average the teachers are scoring a raw score of 91. The conversion table in Figure 7 shows a raw score of 91 is akin to -0.42 logits on the SACS ruler (variable map). And, -0.42 logits, when looking at the variable map in Figure 6 aligns directly with item AR2. This means that, on average, the teachers in this school have a 50% chance of agreeing that “K-12 schools in this region understand the importance of STEM education.” All items below this point are more easily endorsed by teachers in this school and focus on the general importance of STEM education for students. So the next step with teachers in this school, if we want to move them up the SACS ruler, would be to provide them with online and local STEM educational resources and potential opportunities for teachers and students—since this is the focus of the next grouping of items. If utilized in this fashion, the SACS can be used as a road map for systematically planning and implementing STEM reform efforts in schools and communities.

CONCLUSIONS

Possessing the ability to use survey results in a meaningful way is of great importance in social science research. However, it is not reasonable to claim reliable and informed judgments about survey research results can be made from tools that have not been shown to be empirically sound. In our survey evaluation study, Rasch analyses allowed us to modify our instrument based on empirically driven recommendations to delete some items, modify others, and change our scale to elicit a unidimensional measurement. These survey scale and item changes further helped to produce an instrument that optimized results obtained from data collected and minimized measurement error. In survey research, this has great value since we often only have one opportunity to communicate with our participants through the gathering of their feelings, beliefs, or attitudes in response to our surveys. Thus, it is imperative that all, researchers and survey respondents, have a shared understanding of the survey items and scale meaning to interpret results accurately. Rather than assuming this to be the case, Rasch measurement allowed us to assess this empirically and advance the field of science education research by developing a needed affective measure of STEM attitudes and support by various community partners. With the growing number of funded STEM education initiatives nationally, this instrument fills a much-needed niche as a tool to evaluate the impact of STEM investments on STEM community awareness.

Furthermore, this study, coupled with three prior works in *Science Education*, demonstrate the importance and application of Rasch measurement in a variety of contexts. Science education researchers now have practical examples of Rasch measurement studies conducted using dichotomous data (Rasch, 1960) when examining a state-level multiple-choice test in science (Boone & Scantlebury, 2006), the Rasch partial credit model (Masters, 1982) for developing an assessment of students’ socioscientific decision-making strategies (Eggert & Bögenholz, 2009), the Rasch rating scale model (Andrich, 1978) used for assessing a previously made survey of self-efficacy (Boone et al., 2010), and now the Rasch rating scale model (Andrich, 1978) used for developing a new survey (current study). Perhaps these works together will act as a primer for science education researchers interested in measurement issues and provide a meaningful science context for understanding these issues better before taking on more complete Rasch textbooks (e.g., Bond & Fox, 2007).

APPENDIX A
Comparison of Pilot 1 to Pilot 2 Item Measures

Final Survey Item	Initial Survey Item	Initial to Final Item Action	Final Survey Item Measure (SE)	Initial Survey Item Measure (SE)	Initial to Final Item Comparison
Industry Engagement in STEM Education IE1. I believe it is important for area businesses to be involved in STEM partnership(s) with K-12 schools in my region	Same as final survey item	N/A	-1.53 (0.07)	-1.61 (0.18)	Unchanged within $\pm 2 SE$
IE2. I have had businesses/community funded STEM education programs or events in my school or school district	Same as final survey item	N/A	0.16 (0.06)	-0.17 (0.20)	Unchanged within $\pm 2 SE$
IE3. I have had community/business volunteers for STEM education programs or events in my school or school district	Same as final survey item	N/A	-0.06 (0.06)	-0.36 (0.20)	Unchanged within $\pm 2 SE$
IE4. I have had community/business guest speakers in my school or school district	Same as final survey item	N/A	-0.68 (0.07)	-0.91 (0.19)	Unchanged within $\pm 2 SE$
N/A	IE5. I have been involved in coteaching STEM lessons with community/business members.	Removed from initial to final survey due to misfit	N/A	0.16 (0.18)	N/A

Continued

Final Survey Item	Initial Survey Item	Initial to Final Item Action	Final Survey Item Measure (SE)	Initial Survey Item Measure (SE)	Initial to Final Item Comparison
N/A	IE6. I have been involved in STEM curriculum planning with community/business stakeholders	Removed from initial to final survey due to misfit.	N/A	0.00 (0.18)	N/A
IE5. There are opportunities for K-12 students to complete internships or co-ops in this region	IE7 on initial survey is same item content as IE5 on final survey	Moved up two places in item ordering because of removed items	0.05 (0.06)	0.35 (0.19)	Unchanged within ± 2 SE
IE6. There are organizations interested in providing STEM education opportunities for K-12 students in this region	IE8 on initial survey is same item content as IE6 on final survey	Moved up two places in item ordering because of removed items.	-0.21 (0.07)	-0.56 (0.17)	Unchanged within ± 2 SE
N/A	IE9. There has been an increase in K-12 STEM education opportunities offered by organizations within my region in the last year	Removed from initial to final survey due to redundancy in item content and difficulty	N/A	-0.04 (0.18)	N/A
IE7. Overall, there has been an increase in K-12 STEM education opportunities for students in the region in the last year	IE10 on initial survey is same item content as IE7 on final survey	Moved up three places in item ordering because of removed items	0.06 (0.06)	-0.08 (0.18)	Unchanged within ± 2 SE

Continued

Final Survey Item	Initial Survey Item	Initial to Final Item Action	Final Survey Item Measure (SE)	Initial Survey Item Measure (SE)	Initial to Final Item Comparison
IE8. I have worked closely with community/business organization members in my role as an educator	IE11 on initial survey is same item content as IE8 on final survey	Moved up three places in item ordering because of removed items	1.12 (0.06)	0.10 (0.21)	Item significantly more challenging to endorse on revised survey
N/A	IE12. It is important for educators to build relationships with members of the outside community	Removed from initial to final survey due to redundancy in item content and difficulty	N/A	-2.10 (0.20)	N/A
STEM Awareness and Resources					
AR1. My school district understands the importance of STEM education	Same as final survey item	N/A	-1.05 (0.07)	-0.91 (0.20)	Unchanged within ± 2 SE
AR2. The K-12 schools in this region understand the importance of STEM education	Same as final survey item	N/A	-0.28 (0.07)	-0.18 (0.18)	Unchanged within ± 2 SE
AR3. Parents in this region understand the importance of STEM education	Same as final survey item	N/A	1.18 (0.06)	1.37 (0.22)	Unchanged within ± 2 SE
AR4. More work needs to be completed to spread awareness of STEM education	Same as final survey item	N/A	-1.71 (0.07)	-1.84 (0.19)	Unchanged within ± 2 SE
AR5. STEM skills are integral to student success today	Same as final survey item	N/A	-1.70 (0.07)	-1.62 (0.18)	Unchanged within ± 2 SE
AR6. Increasing the STEM talent pool is necessary for economic vitality	Same as final survey item	N/A	-1.83 (0.07)	-1.95 (0.19)	Unchanged within ± 2 SE

Continued

Final Survey Item	Initial Survey Item	Initial to Final Item Action	Final Survey Item Measure (SE)	Initial Survey Item Measure (SE)	Initial to Final Item Comparison
AR7. Students with postsecondary education are more likely to secure a career in a STEM field	N/A	New item created for final survey	-1.50 (0.07)	N/A	N/A
N/A	AR7. Students can be successful in college without a firm understanding of STEM subjects	Removed from initial to final survey due to misfit.	N/A	1.01 (0.21)	N/A
N/A	AR8. Students must complete a 4-year college degree to secure a career in a STEM field	Removed from initial to final survey due to misfit	N/A	0.97 (0.21)	N/A
AR8. There are colleges and/or universities and/or community colleges that offer scholarships for students to pursue STEM degrees in my region	AR9 on initial survey is same content as AR8 on final survey	Moved up one place in item ordering because of removed/added items	-0.58 (0.07)	-0.21 (0.18)	Unchanged within ± 2 SE
AR9. There are STEM education Web sites available for this region that include activities for K-12 teachers and students	AR10 on initial survey is same content as AR9 on final survey	Moved up one place in item ordering because of removed/added items	-0.17 (0.07)	-0.46 (0.17)	Unchanged within ± 2 SE
AR10. Information on regional STEM career opportunities is available online	AR11 on initial survey is same content as AR10 on final survey	Moved up one place in item ordering because of removed/added items	-0.05 (0.06)	0.21 (0.18)	Unchanged within ± 2 SE

Continued

Final Survey Item	Initial Survey Item	Initial to Final Item Action	Final Survey Item Measure (SE)	Initial Survey Item Measure (SE)	Initial to Final Item Comparison
AR11. Local organizations recruit STEM talent online	AR12 on initial survey is same content as AR11 on final survey	Moved up one place in item ordering because of removed/ added items	0.56 (0.06)	0.97 (0.21)	Unchanged within $\pm 2 SE$
AR12. Information related to STEM opportunities in my region is available online	AR13 on initial survey is same content as AR12 on final survey	Moved up one place in item ordering because of removed/ added items	0.19 (0.06)	-0.70 (0.17)	Item significantly more challenging to endorse on revised survey
AR13. There are other STEM online tools available to this region	AR14 on initial survey is same content as AR13 on final survey	Moved up one place in item ordering because of removed/ added items	0.15 (0.06)	0.04 (0.18)	Unchanged within $\pm 2 SE$
PR1. Students in this region are prepared by K-12 schools to be successful in postsecondary study (2- or 4-year colleges or universities and technical programs)	Same as final survey item	N/A	0.71 (0.06)	0.29 (0.19)	Unchanged within $\pm 2 SE$
PR2. Students in this region are knowledgeable about the STEM careers that will be in high demand when they graduate	Same as final survey item	N/A	1.13 (0.06)	1.31 (0.22)	Unchanged within $\pm 2 SE$
PR3. The K-12 public schools in this region effectively teach students STEM knowledge and skills	Same as final survey item	N/A	.098 (0.06)	1.02 (0.21)	Unchanged within $\pm 2 SE$

Continued

Final Survey Item	Initial Survey Item	Initial to Final Item Action	Final Survey Item Measure (SE)	Initial Survey Item Measure (SE)	Initial to Final Item Comparison
PR4. The state standardized tests used in this region's K-12 schools adequately assess STEM knowledge and skills	Same as final survey item	N/A	1.50 (0.06)	1.41 (0.22)	Unchanged within ± 2 SE
PR5. The K-12 schools in this region prepare students who are critical thinkers and problem solvers	Same as final survey item	N/A	1.33 (0.06)	0.80 (0.21)	Unchanged within ± 2 SE
PR6. Community partners (e.g., business and high education) in this region are engaged in making K-12 STEM education more relevant through providing real-world connections in this	Same as final survey item	N/A	0.66 (0.06)	0.75 (0.21)	Unchanged within ± 2 SE
Regional STEM Careers and Workforce CW1. There are businesses and industries that provide signing bonuses and/or incentives for workers choosing a STEM career in the region	Same as final survey item	N/A	1.09 (0.06)	1.29 (0.22)	Unchanged within ± 2 SE
CW2. Organizations have experiences an increase in the number of STEM positions available in the last year in this region	Same as final survey item	N/A	0.76 (0.06)	1.19 (0.22)	Unchanged within ± 2 SE
CW3. Organizations have been able to fill all STEM-related positions within the last year in this region	Same as final survey item	N/A	1.31 (0.06)	1.48 (0.23)	Unchanged within ± 2 SE

Continued

Final Survey Item	Initial Survey Item	Initial to Final Item Action	Final Survey Item Measure (SE)	Initial Survey Item Measure (SE)	Initial to Final Item Comparison
CW4. Organizations have experienced an increase in the number of women and minorities in STEM positions in the last year in this region	Same as final survey item	N/A	1.16 (0.06)	1.45 (0.23)	Unchanged within ± 2 SE
CW5. Organizations have been able to fill STEM-related positions with local STEM talent	Same as final survey item	N/A	1.39 (0.06)	1.00 (0.21)	Unchanged within ± 2 SE
N/A	CW6. K-12 educators and administrators in this region are aware of the workforce needs of area employers	Removed from initial to final survey due to misfit	N/A	0.47 (0.23)	N/A
CW6. It is important for businesses in this region to be able to recruit skilled workers locally	CW7 on initial survey is same item content as CW6 on final survey	Moved up one place in item ordering because of removed items	-1.65 (0.07)	-1.69 (0.19)	Unchanged within ± 2 SE
CW7. All students should receive information about careers that are expected to be in demand in this region when they graduate from K-12 schools and postsecondary institutions	CW8 on initial survey is same item content as CW7 on final survey	Moved up one place in item ordering because of removed items.	-1.70 (0.07)	-1.73 (0.19)	Unchanged within ± 2 SE
CW8. All K-12 schools in this region teach the STEM skills and knowledge appropriate for jobs that will be available in the region	CW9 on initial survey is same item content as CW8 on final survey	Moved up one place in item ordering because of removed items.	0.85 (.06)	1.00 (0.22)	Unchanged within ± 2 SE

Continued

Final Survey Item	Initial Survey Item	Initial to Final Item Action	Final Survey Item Measure (SE)	Initial Survey Item Measure (SE)	Initial to Final Item Comparison
CW9. All K-12 students should have access to STEM education N/A	N/A	New item created for final survey	-1.35 (0.07)	N/A	N/A
	CW10. STEM education is for all students	Removed from initial to final survey due to misfit	N/A	-0.88 (0.18)	N/A
CW10. Career-oriented education is for all students	CW11 on initial survey is same item content as CW10 on final survey	Moved up one place in item ordering because of removed/added items	-0.62 (0.07)	-1.07 (0.18)	Unchanged within ± 2 SE
CW11. Preparing students for careers in STEM is a top priority for schools in this region	CW12 on initial survey is same item content as CW11 on final survey	Moved up one place in item ordering because of removed/added items	0.89 (0.06)	0.87 (0.21)	Unchanged within ± 2 SE
CW12. Stakeholders within community/business organizations have STEM skills and knowledge that could be an asset to K-12 schools in this region	CW13 on initial survey is same item content as CW12 on final survey	Moved up one place in item ordering because of removed/added items	-0.54 (0.07)	-0.43 (0.18)	Unchanged within ± 2 SE

APPENDIX B**Final K-12 STEM Awareness and Community Support Survey****Industry Engagement in STEM Education (IE)**

1. I believe it is important for area businesses to be involved in STEM partnership(s) with K-12 schools in my region.
2. I have had business/community funded STEM education programs or events in my school or school district.
3. I have had community/business volunteers for STEM education programs or events in my school or district.
4. I have had community/business guest speakers in my school or school district.
5. There are opportunities for K-12 students to complete internships or coops in the region.
6. There are organizations interested in providing STEM education opportunities for K-12 students in this region.
7. Overall, there has been an increase in K-12 STEM education opportunities for students in the region in the last year.
8. I have worked closely with community/business organization members in my role as an educator.

STEM Awareness and Resources (AR)

1. My school district understands the importance of STEM education.
2. The schools in this region understand the importance of STEM education.
3. Parents in this region understand the importance of STEM education.
4. More work needs to be completed to spread awareness of STEM education.
5. STEM skills are integral to student success today.
6. Increasing the STEM talent pool is necessary for economic vitality.
7. Students with postsecondary education are more likely to secure a career in a STEM field.
8. There are colleges and/or universities and/or community colleges that offer scholarships for students to pursue STEM degrees in my region.
9. There are STEM education Web sites available for this region that include activities for teachers and students.
10. Information on regional STEM career opportunities is available online.
11. Local organizations recruit STEM talent online.
12. Information related to STEM opportunities in my region is available online.
13. There are other STEM online tools available to this region.

Preparation of Students for Success in College & Careers (PR)

1. Students in this region are prepared by K-12 schools to be successful in postsecondary study (2- or 4-year colleges or universities and technical programs).
2. Students in this region are knowledgeable about the STEM careers that will be in high demand when they graduate.
3. The K-12 public schools in this region effectively teach students STEM knowledge and skills.
4. The state standardized tests used in this region's K-12 schools adequately assess STEM knowledge and skills.
5. The K-12 schools in this region prepare students who are critical thinkers and problem solvers.

Continued

6. Community partners (e.g., business and higher education) are engaged in making K-12 STEM education more relevant through providing real-world connections in this region.

Regional STEM Careers and Workforce (CW)

1. There are businesses and industries that provide signing bonuses and/or incentives for workers choosing a STEM career in the region.
2. Organizations have experienced an increase in the number of STEM positions available in the last year in this region.
3. Organizations have been able to fill all STEM-related positions within the last year in this region.
4. Organizations have experienced an increase in the number of women and minorities in STEM positions in the last year in this region.
5. Organizations have been able to fill STEM-related positions with local STEM talent.
6. It is important for businesses in this region to be able to recruit skilled workers locally.
7. All students should receive information about careers that are expected to be in demand in this region when they graduate from K-12 schools and postsecondary institutions.
8. All K-12 schools in this region teach the STEM skills and knowledge appropriate for jobs that will be available in the region.
9. All K-12 students should have access to STEM education.
10. Career-oriented education is for all students.
11. Preparing students for careers in STEM is a top priority for schools in this region.
12. Stakeholders within community/business organizations have STEM skills and knowledge that could be an asset to K-12 schools in this region.

Note: All items are on a 1–4 point Likert-scale with 1 = *strongly disagree*, 2 = *disagree*, 3 = *agree*, 4 = *strongly agree*. No items are reverse coded.

REFERENCES

- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561–573.
- Bode, R., & Wright, B. (1999). Rasch measurement in higher education. In J. Smart & W. Tierney (Eds.), *Higher education handbook of theory and research* (Vol. XIV, p. 287–316).
- Bond, T., & Fox, C. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.
- Boone, W. J., & Scantlebury, K. (2006). The role of Rasch analysis when conducting science education research utilizing multiple-choice tests. *Science Education*, 90(2), 253–269.
- Boone, W. J., Townsend, J. S., & Starver, J. (2010). Using Rasch theory to guide the practice of survey development and survey data analysis in science education and to inform science reform efforts: An exemplar utilizing STEBI self-efficacy data. *Science Education*, 95(2), 258–280.
- Breiner, J., Harkness, M., Johnson, C. C., & Koehler, C. (2012). What Is STEM? A discussion about conceptions of STEM in education and partnerships. *School Science and Mathematics*, 112(1), 3–11.
- Bureau of Labor Statistics. (2008). *Employment projections: 2008–2018 summary*. Retrieved from <http://www.bls.gov/news.release/ecopro.nr0.htm>.
- Czerniak, C. M. (2009). Grand challenges and great opportunities in science education: Is the glass half full or half empty? NARST Presidential Speech. *E-NARST News*, 52(2), 3–8. Retrieved from http://www.narst.org/news/e-narstnews_july2009.pdf.
- Duncan, P., Bode, R., Lai, S., & Perera, S. (2003). Rasch analysis of a new stroke-specific outcome scale: The stroke impact scale. *Archives of Physical Medicine and Rehabilitation*, 84, 950–963.
- Eggert, S., & Bogeholz, S. (2009). Students' use of decision-making strategies with regard to socioscientific issues: An application of the Rasch partial credit model. *Science Education*, 94(2), 230–258.

- Enochs, L. G., & Riggs, I. M. (1990). Further development of an elementary science teaching self-efficacy belief instrument: A perspective elementary scale. *School Science and Mathematics*, 90, 694–706.
- Hambleton, R. K. (2000). Emergence of item response modeling in instrument development and data analysis. *Medical Care*, 38(9), 60–65.
- Johnson, C. C. (2012). Implementation of STEM education policy: Challenges, progress, and lessons learned. *School Science and Mathematics*, 112(1), 45–55.
- Linacre, J. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, 31, 85–106.
- Linacre, J. (2012). *WINSTEPS Rasch measurement*. Chicago: MESA Press.
- Liu, X. (2010). Using and developing measurement instruments in science education: A Rasch modeling approach. Charlotte, NC: Information Age.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149–174.
- Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Danmarks Paedagogiske Institut.
- Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests (expanded ed.). Chicago: MESA Press.
- Reise, S., & Henson, J. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment*, 81, 93–103.
- Shirley, D. (2009). Community organizing and organizing change: A reconnaissance. *Journal of Educational Change*, 10, 229–237.
- Smith, E., Conrad, K., Chang, K., & Piazza, J. (2002). An introduction to Rasch measurement for scale development and person assessment. *Journal of Nursing Measurement*, 10, 189–206.
- Smith, R. (1996). A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling*, 3, 25–40.
- Waugh, R., & Chapman, E. (2005). An analysis of dimensionality using factor analysis (true-score theory) and Rasch measurement: What is the difference? Which method is better? *Journal of Applied Measurement*, 6, 80–99.
- Wilson, M., & Draney, K. (2002). A technique for setting standards and maintaining them over time. In S. Nishisato, Y. Baba, H. Bozdogan, & K. Kanefugi (Eds.), *Measurement and multivariate analysis* (Proceedings of the International Conference on Measurement and multivariate Analysis, Banff, Canada, May 12–14, 2000; pp 325–332). Tokyo: Springer-Verlag.
- Wright, B. D. (1996). Comparing Rasch measurement and factor analysis. *Structural Equation Modeling*, 3, 3–24.
- Wright, B. D., & Linacre, J. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8(3), 370.
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESA Press.
- Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago: MESA Press.
- Zmuda, A., Kuklis, R., & Kline, E. (2004). *Transforming schools: Creating a culture of continuous improvement*. Alexandria, VA: Association for Supervision and Curriculum Development.

Scientific Practices in Elementary Classrooms: Third-Grade Students' Scientific Explanations for Seed Structure and Function

LAURA ZANGORI,¹ CORY T. FORBES²

¹*Department of Teaching, Learning, and Teacher Education, College of Education and Human Sciences, and* ²*School of Natural Resources, University of Nebraska–Lincoln, Lincoln, NE 68583, USA*

Received 19 May 2013; accepted 28 March 2014

DOI 10.1002/sce.21121

Published online 14 May 2014 in Wiley Online Library (wileyonlinelibrary.com).

ABSTRACT: Elementary science standards emphasize that students should develop conceptual understanding of the characteristics and life cycles of plants (National Research Council, 2012), yet few studies have focused on early learners' reasoning about seed structure and function. The purpose of this study is twofold: to (a) examine third-grade students' formulation of explanations about seed structure and function within the context of a commercially published science unit and (b) examine their teachers' ideas about and instructional practices to support students' formulation of scientific explanations. Data, collected around a long-term plant investigation, included classroom observations, teacher interviews, and students' written artifacts. Study findings suggest a link between the teachers' ideas about scientific explanations, their instructional scaffolding, and students' written explanations. Teachers who emphasized a single "correct explanation" rarely supported their students' explanation-construction, either through discourse or writing. However, one teacher emphasized the importance of each student generating his/her own explanation and more frequently supported students to do so in the classroom. The evidentiary basis of her students' written explanations was found to be much stronger than those from students in the other two classrooms. Overall, these findings indicate that teachers' conceptions about

Correspondence to: Laura Zangori; e-mail: laura.zangori@huskers.unl.edu

An earlier version of this paper was presented at the 2013 meeting of the National Association for Research in Science Teaching in Rio Grande, Puerto Rico.

scientific explanations are crucial to their instructional practices, which may in turn impact students' explanation-construction. © 2014 Wiley Periodicals, Inc. *Sci Ed* 98:614–639, 2014

INTRODUCTION

Plant growth and development is a foundational scientific concept that spans K-12 curriculum standards (National Research Council [NRC], 2000, 2012). Elementary science standards specifically emphasize that early learners should develop conceptual understanding about characteristics and life cycles of organisms, as well as interactions between organisms and their environment (NRC, 2000, 2012). The focus on plant characteristics and life cycles in the early grades is particularly important because some evidence suggests that as children develop, their ability to notice plants, their assumptions about the importance of plants, and their interest in plants deteriorates (e.g., *plant blindness*, Wandersee & Schussler, 1999). The conceptual understanding students develop about plants in the elementary grades therefore serves as a foundation for later science learning (Duschl, Schweingruber, & Schouse, 2007). Although education research has predominantly focused on students' alternate conceptions about photosynthesis (e.g., Canal, 1999), few studies have focused on early learners' scientific reasoning about plant growth and development (e.g., Beyer & Davis, 2008; Jewel, 2002; Metz, 2008), particularly early learners' understanding of seed structure and function.

Scientific explanation-construction is a crucial scientific practice that helps facilitate students' conceptual development (Duschl et al., 2007; NRC, 2012). It is defined by opportunities for students to connect observable cause and effect with an underlying unseen mechanism (Braaten & Windschitl, 2011; NRC, 2012). While engaging in this process, students are able to address an investigation question empirically and examine their existing knowledge in light of new knowledge (NRC, 2000). A growing literature base has documented elementary students' abilities to engage in explanation-construction (Herrenkohl, Palincsar, DeWater, & Kawasaki, 1999; Mason, 2001; Metz, 2008; Ryu & Sandoval, 2012; Zuzovsky & Tamir, 1999). However, for elementary students to formulate scientifically accepted, mechanism-based explanations effectively, scaffolding in multiple forms is required in knowledge-rich learning environments, providing students with opportunities to challenge their preexisting naïve explanations and construct new knowledge (Mason, 2001; Ryu & Sandoval, 2012). Yet, explanation-construction is frequently deemphasized in elementary science learning environments (Forbes, Biggers, & Zangori, 2013; Metz, 2008; Zangori, Forbes, & Biggers, 2013). Like those designed for middle-school and secondary science (Beyer, Delgado, Davis, & Krajcik, 2009; Kesidou & Rosemann, 2002), widely available science curriculum materials used in the elementary classrooms tend to not prioritize opportunities for students to propose a mechanism for observed cause and effect (Biggers, Forbes, & Zangori, 2013; Kuhn, 2009; Metz, 2004; Zangori et al., 2013). Furthermore, some evidence suggests that even when explanation-construction is emphasized in curriculum materials, teachers may not enact them as intended (e.g., Beyer & Davis, 2008; Metz, 2009), a finding consistent with theoretical perspectives on the teacher-curriculum relationship (Davis & Krajcik, 2005; Remillard, 2005).

More work is needed to understand how elementary students can be supported to formulate scientific explanations, particularly about topics such as seed structure and function where students exhibit a variety of alternate conceptions. Here, we examine explanation-construction within the context of a long-term investigation about plants in three third-grade classrooms. We ask the following research questions:

1. How do third-grade students formulate written scientific explanations about seed structure and function?
2. In what ways and why do third-grade teachers provide instructional support for students' formulation of scientific explanations about seed structure and function?

BACKGROUND AND CONCEPTUAL FRAMEWORK

Scientific Explanations in the Elementary Classroom

The recently released *Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* (NRC, 2012) defines scientific explanations as

... accounts that link scientific theory with specific observations or phenomena ... they explain observed relationships between variables and describe the mechanisms that support cause and effect inferences about them. (p. 67)

Consistent with this definition, we define a scientific explanation as identification of a *mechanism* that underlies observable *cause* and *effects* or, as Zimmerman (2007) describes, the “process by which a cause can bring about an effect” (p. 184). This perspective on explanation-construction in science aligns with current science education literature (e.g., Braaten & Windschitl, 2011; Duschl et al., 2007) and those within the philosophy of science (Salmon, 1998). Observable causes and effects represent the “what” of natural phenomena, whereas the mechanism is the often unobservable causal force for “how” and “why” a phenomenon occurred, derived from experience in the classroom, that is within the norms of discourse and the production of artifacts (Braaten & Windschitl, 2011; Herrenkohl et al., 1999; Schauble, 1996).

Students, as well as adults, do not develop the capacity for scientific reasoning without frequent sensemaking opportunities and ongoing guidance and support (Herrenkohl et al., 1999; Kuhn, 2009; Ryu & Sandoval, 2012; Schauble, 1996). However, most widely available, off-the-shelf science curriculum materials for kindergarten through third grade are grounded in cognitive developmental theory that assumes early learners are “concrete” thinkers and therefore foreground categorization, classification, and serration of decontextualized science concepts (Metz, 2004). When opportunities for sense making are absent, early learners will draw upon naïve causal mechanisms based on their prior experiences and understandings. These preexisting mechanism “libraries” (Schauble, 1996, p. 112) are frequently different from the scientifically accepted mechanisms, as they are typically developed in the absence of domain-specific knowledge about the phenomenon or after being provided support by an experienced other that can help students notice what they may not see on their own.

The evidence for students' preexisting mechanism libraries is apparent in almost three decades of research on plant growth and development. Alternate conceptions about seed growth and development are prevalent in elementary school (Barman, Stein, McNair, & Barman, 2006; Canal, 1999; Jewel, 2002; Patrick & Tunnicliffe, 2011) and remain into adulthood (Wandersee & Schussler, 1999). Students articulate naïve conceptions about the purpose of the seed, where the seed comes from, what is contained inside the seed, what the dormant seed needs to grow, and whether it is living or nonliving. For example, children do not consider seeds living things until they are planted and watered. This alternative conception may be attributed to children's prior experiences with planting seeds underground. When sprouts appear above ground, the only mechanism available to them is

somewhat akin to “magic” as they do not have access to alternative causal mechanism for their observations of seed germination.

When students encounter seed growth and development in the classroom, they may integrate new information into these preexisting explanations, but if the new information is not well understood and connected into a coherent causal story, they may develop a new alternate conception. This new alternate conception may contain *some* elements of the new knowledge, but it will most likely be incorporated into the existing knowledge so that the mechanism will contain elements of naïve understandings attached to scientific principles (Duschl et al., 2007). To provide student opportunities for active sense making (Mason, 2001; Ryu & Sandoval, 2012), there are several epistemic commitments of which scientific explanation-construction should be composed. These include constructing explanations that (a) answer an investigation question, (b) are based on data and evidence that support answering the investigation question, (c) provide opportunities for new understanding; and (d) build on preexisting ideas (Biggers et al., 2013; Forbes et al., 2013; NRC, 2000, 2012; Zangori et al., 2013). When elementary students are provided opportunities to generate scientific explanations in learning environments that have been carefully crafted to address these epistemic commitments, past research has shown that early learners are able to articulate cause, effect, and mechanism for phenomena such as floating and sinking (Hardy, Jonen, Möller, & Stern, 2006; Herrenkohl et al., 1999), force and motion (Hapgood, Magnusson, & Palincsar, 2004), animal behavior (Metz, 2008), and plants (Mason, 2001; Metz, 2008).

Supporting Students' Explanation-Construction

Recent research suggests that elementary teachers follow their science curriculum materials closely (Biggers et al., 2013; Forbes et al., 2013; Zangori et al., 2013), even though commercially produced curricular tools exhibit a variety of limitations (Beyer et al., 2009; Kesidou & Rosemann, 2002; Metz, 2004, 2008; Patrick & Tunnicliffe, 2011; Schussler, 2008). Specifically for life sciences, topics are fragmented within the science curriculum, making it difficult for students and their teachers to bring together the big ideas with which they engage and/or for teachers to support students in anchoring new knowledge (Metz, 2004; Stern & Rosemann, 2004). While there has been no comprehensive review of elementary science curriculum materials to date (Kesidou & Rosemann, 2002), Schussler (2008) reviewed 69 botanical trade books used frequently in conjunction with elementary curricular units and identified five consistently observed sources of error and omission. These included no mention of fruits or seeds in the plant lifecycle or, if fruits and seeds were included, no explanations for what their function is or how they “appeared” within the life cycle.

To challenge students' alternative conceptions and provide opportunities to develop new explanations, students require scaffolding in multiple forms, both curricular and instructional (Hardy et al., 2006; Herrenkohl et al., 1999). Whole-class and individual discussions in which teachers support students in making observations and discussing competing theories are effective in providing student opportunities in connecting cause, effect, and mechanism and facilitating conceptual change. As Hardy and colleagues (2008) found, as well as others (e.g., Hapgood et al., 2004; Herrenkohl et al., 1999; Mason, 2001; Metz, 2008), the teacher and curriculum working together synergistically (Tabak, 2004) are major factors in effectively designed science learning environments that provide opportunities for generating scientific explanations and foreground conceptual change.

Explanation-construction may take a variety of more student- or teacher-directed forms in the classroom depending on the teacher's knowledge about generating scientific

explanations (Beyer & Davis, 2008; Biggers et al., 2013; Forbes et al., 2013; Horwood, 1988; Metz, 2009; Zangori et al., 2013; Zuzovsky & Tamir, 1999). Teachers who locate explanatory authority in the curriculum materials (and see their responsibility as conveying this information to their students) will tend to focus their instruction on providing students with a single “correct explanation” (Zuzovsky & Tamir, 1999 p. 1120). Within this perspective on explanation-construction, there is no need for students to examine alternative explanations. Teachers who see their responsibility as encouraging students to examine explanations for adequacy and evaluate them against other possible explanations use a “structure of science” (Zuzovsky & Tamir, 1999, p. 1120) model. From this perspective, alternative explanations exist; however, the teacher defines her role to support students in evaluating explanations until a satisfactory explanation is constructed. Finally, teachers who support their students to connect cause, effect, and mechanism but consider all explanations valid regardless of their nature or acceptability of the mechanism are characterized by the “self as explainer” model (Zuzovsky & Tamir, 1999, p. 1120). From this perspective, a teacher may make no attempts to support students to evaluate their explanations in light of other explanations.

However, the ways in which different teachers enact the same curriculum materials are dependent on a number of factors, including their knowledge about their students’ capabilities, their ideas about scientific explanations, and the ways in which they establish discourse in the classroom (Beyer & Davis, 2008; Biggers et al., 2013; Forbes et al., 2013; Enyedy & Goldberg, 2004; Forbes & Davis, 2010; Metz, 2009; Zangori et al., 2013). These factors affect whether they engage students’ in explanation-construction and, if so, how and to what extent. While a research base for elementary students’ scientific reasoning exists, little is known about the ways in which elementary teachers use elementary science curriculum materials to provide opportunities for students to construct scientific explanations and the subsequent success students may have in generating mechanism-based explanations, particularly about seed structure and function.

METHODS

In this concurrent mixed methods study (Creswell & Plano Clark, 2011), we examined how students ($n = 59$) and teachers ($n = 3$) in three 3rd-grade classrooms engaged in explanation-construction during enactment of the 8-week Full Option Science System (FOSS) elementary science unit on plant growth and development titled *Structures of Life* (FOSS, 2005). We used both quantitative and qualitative methods to analyze students’ written artifacts. We used qualitative methods to analyze video-recorded classroom observations for evidence of teachers’ support for students’ formulation of scientific explanations. We used interview data to provide insight into how and why teachers engaged their students in scientific explanation-construction.

Study Context and Participants

The three elementary teachers in this study—Grace, Emily, and Janet—were participants in a 3-year professional development program designed to support elementary teachers in a large, urban school district to learn to evaluate and adapt newly introduced, kit-based elementary science curriculum materials to better engage students in scientific practices (Biggers et al., 2013; Forbes et al., 2013; Zangori et al., 2012, 2013). The project involved 44 in-service elementary teachers from the partner district as well as four surrounding districts within a single midwestern state. Participation in the project was voluntary, and all participants were compensated for their involvement. Three teachers were purposefully

TABLE 1
Summary Profiles of Study Teachers and Classrooms

Demographics	Grace	Emily	Janet
Graduate education	MS in reading literacy	MS in teaching and leadership	None
Years of teaching experience	34	16	8
School	Eastwood	Eastwood	Northwood
Class size	22	20	17
Average lesson length (minutes)	84	55	57

Note: Schools and teachers are identified by pseudonyms.

sampled (Creswell & Plano Clark, 2011) for this study because each (a) was a third-grade teacher, (b) taught the FOSS *Structures of Life* unit during the same 3-month period of the school year, (c) was using this curricular unit for only the second time after its adoption by the district, (d) taught in similar school settings, and (e) was in the postinduction phase of her career (see Table 1).

Each of these three teachers reported taking a standard science methods course during his or her undergraduate teacher education program. In addition, each teacher readily participated in district-provided science workshops throughout her tenure in this district. These workshops, led by the district science coordinator, focused on effective uses of science notebooks, the five essential features of inquiry (NRC, 2000), and science content knowledge. Furthermore, Grace and Emily had both taken additional graduate science courses geared for in-service elementary teachers, which included summer research experiences. Grace and Janet also served as the science coordinators for their respective schools during this study.

The FOSS *Structures of Life* curricular unit (FOSS, 2005) focuses on organismal structure and function and is widely used across this midwestern state, as well as nationally. The FOSS curriculum series, developed at the Lawrence Hall of Science, was introduced in 1988 and is used widely in the United States. It has a scope and sequence ranging from kindergarten through eighth grade, covering topics in the life sciences, as well as physical and earth sciences. FOSS curricular units are grounded in Piagetian perspectives on learning, emphasizing cognitive stages of development (Lowery, 1998; Metz, 2004):

The FOSS program is guided by research on human cognitive development. The activities and intellectual demands are matched to the ways students think at different times in their lives. . . . In their early elementary years, students learn science best from direct experiences in which they observe, describe, sort, and organize objects, organisms, materials, and simple systems. . . . Upper elementary students construct more advanced concepts by classifying, testing, experimenting, and determining cause-and-effect relationships among objects, organisms, and systems. (FOSS, 2005, p. 4)

According to this developmental perspective on learning, K-4 students are at a “concrete” operational level and reasoning abilities have not yet developed (Lowery, 1998). Science experiences for early learners therefore emphasize hands-on experiences where they look for patterns, classify, seriate, and describe their investigations. Research has shown how teachers supplement these materials to provide additional opportunities for scientific sense making (Biggers et al., 2013; Metz, 2004; Ryu & Sandoval, 2012; Zangori et al., 2013).

The first two of the four unit investigations focus on plants. This study occurred during Investigation 1, titled *Origin of Seeds* (Table 2). Students are initially introduced to the

TABLE 2
Overview of FOSS Structure of Life Investigation 1 (Origin of Seeds) Lessons

Order Enacted	Location Within Curriculum Materials	Curriculum Investigation Question	Curriculum Description	Curriculum Data Analysis	Curriculum Explanation
1 ^a	Part 1: Seed search	“Where do seeds come from?” ^c “Where are seeds found on plants?” ^c	Dissection of a bean pod to locate, count, and compare and contrast seeds	Count and graph seeds in a bean pod	N/A
2 ^b	Part 1: Seed search	“Where do seeds come from?” ^c “Where are seeds found on plants?” ^c	Dissection of various fruits to locate, count, and compare and contrast seeds	Count and sort seeds from different fruits	None
3	Part 2: The sprouting seed	“Can a seed grow without soil?” ^c “What effect does water have on seeds?” ^c	Hydroponic investigation of seed growth	Compare and contrast seed changes from seeds not in water to seeds in hydroponic growth environments	“Different kinds of fruits have different kinds and numbers of seeds.” ^d

^aLesson 1 ended at the suggested breakpoint within the curriculum materials.

^bLesson 2 is the remainder and conclusion of Lesson 1 after the breakpoint.

^cFOSS (2005, Investigation 1, p. 2).

^dFOSS (2005, Assessment, p. 6).

phenomena (e.g., fruit) with a short discussion and introduction of new vocabulary terms (e.g., properties). Students then engage in a hands-on investigation including observations, dissection, and data collection. Finally, they classify, organize, and look for patterns and relationships within the data and describe the results of their analysis.

Project teachers participated in a week-long professional development workshop in the summer preceding this study. The workshop focused on all five essential features of inquiry (NRC, 2000), which include opportunities for learners to (a) engage with scientifically oriented questions, (b) give priority to evidence, (c) formulate scientific explanations from collected evidence, (d) evaluate explanations, (e) justify and communicate their scientific explanations. During the workshop, the teachers were provided many opportunities to evaluate elementary science lessons of their own choosing for the ways in which the curriculum materials did, or did not, meet criteria for the feature(s) of inquiry (Zangori et al., 2012). The three teachers in this study utilized time in the workshop to analyze the FOSS *Structures of Life* (2005) unit for features of inquiry and make modifications to unit lessons to address limitations they observed for each feature of inquiry, including explanation-construction. This curriculum planning was in anticipation of enactment of the revised unit in the year in which this study took place.

Data Collection

The data for this study were gathered during the academic year following the summer workshop. First, each teacher was interviewed eight times over the course of the study using semistructured (Patton, 2001) interviews. A formal interview was conducted both at the beginning and end of the year. It was designed to elicit the teachers' conceptions about scientific explanations as well as the ways scientific explanation-construction should be supported in elementary science learning environments. Six additional interviews occurred immediately prior to and following each lesson enactment. These reflective grounded interviews asked teachers to reflect upon their planned and enacted lessons for ways in which they engaged students in scientific explanation-construction. All interview protocols were explicitly aligned with the theoretical framework underlying this study. All interviews ($n = 24$) were conducted by one of the authors either in person or by telephone, audio-recorded, and transcribed verbatim.

We also conducted live observations of each teacher for each of the three investigation lessons, each of which was video-recorded (three teachers \times three lessons each = nine observations). After each observed lesson, we also collected copies of all artifacts and documents created through whole-class discussion and activity (e.g., smartboard files, class-generated data graphed on easel pads, and other class-generated lists). The whole-class artifacts were collected either as video files or as electronic documents and catalogued.

In the year prior to the study, the school district adopted the use of science notebooks at all elementary grades. All three study teachers began implementing the science notebooks at the beginning of the school year in which the study took place (September). We collected and scanned 177 student science notebooks. Each science notebook entry was assigned a unique identification number that associated the student artifacts with lesson observation number and was catalogued. The collected data were therefore hierarchical, nested per student per teacher, and longitudinal, covering three sequential lesson enactments.

Data Analysis

Quantitative Analysis. We used the Practices of Science Observation Protocol (P-SOP; Forbes et al., 2013), a recently developed observation protocol for elementary science, to score both the teachers' lesson plans and video-recorded enacted lessons. The P-SOP is grounded in the five features of inquiry (NRC, 2000). One of the features of inquiry the P-SOP is designed to measure is *formulate explanations about phenomenon of interest that answer investigation question*, a core scientific practice in the elementary grades (Beyer & Davis, 2008; Metz, 2008). In the P-SOP, this scientific practice is measured through four epistemic components: Students should formulate explanations that (a) are based on evidence, (b) answer investigation question, (c) propose new understanding, and (d) build on their existing knowledge (NRC, 2000, 2012).

The P-SOP has been found to be a valid and reliable measure of inquiry in elementary science learning environments (Biggers et al., 2013; Forbes et al., 2013; Zangori et al., 2013). In a previous study, interrater reliability was established through joint scoring of 124 video-recorded elementary science lessons. For the feature *formulate explanations about phenomenon of interest that answer investigation question*, the two scorers' scores accounted for 73% of intrascorer variance with an intraclass correlation coefficient of .79 ($p < .001$). The high Cronbach's α value (.83) also suggests a strong degree of internal reliability for this feature measure.

To evaluate students' written explanations, we adapted the P-SOP to develop a four-part scoring rubric (Table 3). The scoring rubric aligns with the epistemic commitments

TABLE 3
Components of Students' Scientific Explanations

Measure	Level Description	Score
1. Students formulate explanations about phenomenon of interest that are based on evidence	The formulated explanation includes causes of effect or establishes relationships based on empirical evidence.	3
	The formulated explanation includes causes of effect or establishes relationships that are partially supported by evidence.	2
	The formulate explanations for causes of effects or establish relationships that are weakly supported by evidence.	1
	The formulated explanation is not supported by evidence.	0
2. Students formulate explanations about phenomenon of interest that answer investigation question	The formulated explanation fully answers an investigation question.	3
	The formulated explanation partially answers an investigation question.	2
	The formulated explanation weakly answers an investigation question.	1
	The formulated explanation does not answer an investigation question.	0
3. Students formulate explanations about phenomenon of interest that propose new understanding	The formulated explanation illustrates learning: New explanation is different from preexisting explanation and proposes new understanding.	3
	The formulated explanation proposes new understanding about some aspect of the preexisting explanations.	2
	The formulated explanation is similar to and reinforces the preexisting explanation.	1
	The formulated explanation does not propose new understanding.	0
4. Students formulate explanations about phenomenon of interest that build on their existing knowledge	The formulated explanation builds on existing knowledge. There are clear connections between preexisting explanations and new generated explanations.	3
	The formulated explanation is partially based upon preexisting explanations. Some element of the new explanation is based on some element of their preexisting explanation. Other aspects of their preexisting explanations may yet be unresolved.	2
	Some relationship is evident between the preexisting and new explanations, though the former may not ground the latter. Their new and old explanations may exist simultaneously rather than the latter building upon the former.	1
	No relationship is present between preexisting and new explanations	0

of explanation-construction (Hapgood et al., 2004; Hardy et al., 2006; NRC, 2000, 2012; Zangori et al., 2013) and with the four components of explanation-construction identified in the P-SOP (Forbes et al., 2013). To address all four indicators for explanation-construction, we read each student artifact to assess in what ways students were engaging with the four measures. We did not examine the written explanations for explicit statements of each

measure, but rather we examined each student artifact in its entirety to assess if the elements were present and if the written explanation addressed the measures in some way.

For example, to assign a score to the measure *formulate explanations about phenomenon of interest that answer investigation question*, we examined each artifact for an investigation question and then examined the written explanation in light of the investigation question. For each sample, we asked whether the student's explanation answered the investigation question and, if so, in what ways it answered the question. For a student to fully answer the investigation question, all elements of the question had to be addressed in the explanation (Table 3). For an investigation question recorded in an artifact from Grace's classroom, "How do the properties of seeds compare in different kinds of fruits?" (S7, GL2; see Table 4), we examined the written explanation for a discussion of different seeds in different fruits and how they are alike and different. In her explanation, the student included evidence that the different fruits each contained different seed numbers and draws the conclusion that "the bigger the seed the less seeds. The smaller the seed the more seeds." The student addressed property through seed number and seed size and has fully answered the investigation question. This explanation scored a three for the measure *formulate explanations about phenomenon of interest that answer investigation question* (see Table 4).

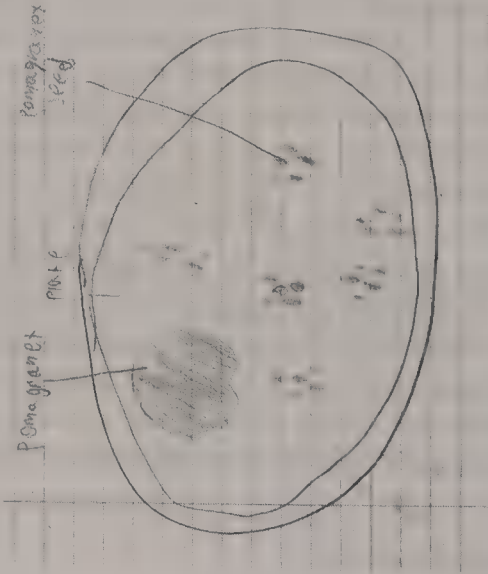
Each explanation was given a score for each measure. The measure scores were summed for each student artifact, providing a composite score ranging from 0 to 12. The first author scored all of the notebook entries. However, prior to scoring all lessons, the authors discussed the rubric and scored together a sample of student artifacts across all three classrooms and lesson to ensure the rubric was adapted from the P-SOP appropriately and the level articulations were appropriate for scoring.

Of the 177 written artifacts collected from students' science notebooks, 29 samples contained no writing pertaining to the lesson. These samples were removed from the data set, leaving 148 student samples available for analysis. The scores for the student writing samples were imported into Statistical Analysis System (SAS; 2013) for multiple regression analysis to examine the relationship between the student writing samples, each teacher's enactment, and each lesson. Our general regression model was $\log y$ (writing samples) = β_0 (error) + β_1 (teacher_{*i*}) + β_2 (lesson_{*i*}) + β_3 (teacher_{*i*} × lesson_{*i*}). We established prior to analyzing the data that the student written artifact, data were not normally distributed due to an abundance of zero composite scores; therefore, the data required a log transformation of the regression outcome (Kleinbaum, Kupper, Muller, & Nizam, 1998). The general regression model includes the relationship between the writing samples, the teacher, and the lesson for each individual teacher. This model was also expanded to a complex model that included all three teachers, all three lessons, and all three interactions where we used "dummy" variables to categorize the teacher and lesson of interest as a 1 and 0 otherwise. We also examined the relationship between the interaction of enactment and lesson with the student writing samples. If an interaction is present, it indicates that neither the enactment nor the lesson individually affected what the student wrote, but rather it is some combination of enactment and lesson that are affecting the student artifacts (Kleinbaum et al., 1998).

Qualitative Analysis

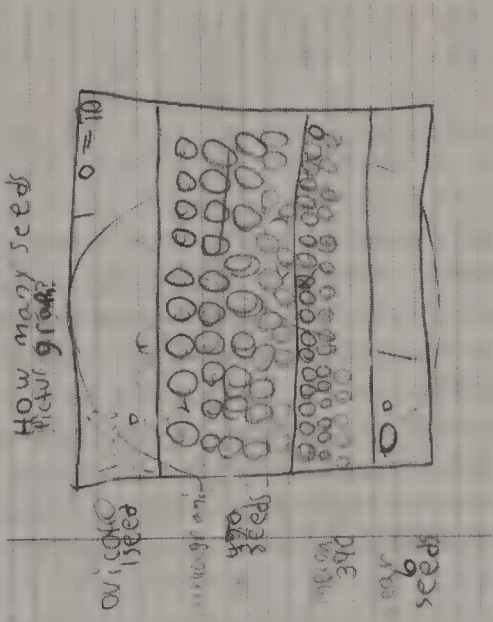
The qualitative portion of this study is a holistic, naturalistic inquiry (Denzin & Lincoln, 2000) using students' writing samples, video-recorded observations, and the teacher interviews. All data were imported into Atlas.ti and coded using classical content analysis (Ryan & Bernard, 2000) for a priori codes of the measures identified in Table 3. Source triangulation was achieved across multiple data sources (e.g., in-depth semistructured interviews, reflective grounded interviews, and lesson observations) through coding

TABLE 4
 Student Scoring Sample

Grace (S7, GL2)		Explanation Rubric Score
Question	How do the properties of the seeds compare in different kinds of fruits.	The formulated explanation fully answers the investigation question as the student addressed property through seed number and seed size in his conclusion. Score: 3
Prediction	The[y] compare different kinds of fruits because the bigger the fruit the bigger the seed. The smaller the fruit the smaller the seed	The formulated explanation illustrates learning. The new explanation of "The bigger the seed the less seeds. The smaller the seed the more seeds." demonstrates that he has changed his thinking from a correlation between fruit and seed size to a correlation between seed size and seed number. Score: 3
Observation		The formulated explanation is partially based upon preexisting explanations. We see that some element of the new explanation (seed size) is based on some element of his preexisting explanation (seed size). However we do not know in this example if the student resolved his issue with fruit size and its relationship to seed size. Score: 2

(Continued)

TABLE 4
Continued

Grace (S7, GL2)		Explanation Rubric Score
Data Analysis		The student refers to the seed count numbers in the explanation. Therefore the explanation established a relationship based on empirical evidence of seed count numbers. Score: 3
Explanation	... The ovicodo [avocado] seed was brown. The pear seeds were black. The pomegranate seeds were a dark pink color. There wer[e] 450 seeds in the pomegranate. The melon had 390. The pear had 6 and the ovicote [avocado] had 1. The bigger the seed the less seeds. The smaller the seeds the more seeds.	Total written explanation score: 11

by two researchers of 10% of the data sources to look for consistency across sources. Interrater reliability among the texts averaged at 90% and, after discussion among the raters, a 100% agreement was reached.

For Research Question 1, each student notebook was examined holistically to identify themes and patterns in the student artifacts within and across teachers. For Research Question 2, we engaged in cross-case analysis of coded data focused on the construction of a multiple-case study of the three teachers. Interview data and observations were examined together to establish themes within and across teachers. The data were triangulated through analysts and sources (Patton, 2001). All qualitative analysis involved an iterative process of data coding, displaying, and verification (Ryan & Bernard, 2000) to provide insight into the students' written explanations, the teachers' instructional practices during lesson enactments, and their conceptions about scientific explanations in elementary science learning environments.

RESULTS

Overall, results from quantitative analysis of the student artifacts indicate that students in Grace's classroom formulated significantly more written explanations than students in Janet or Emily's classrooms. Findings from the qualitative analysis of observations and interviews suggest a link between the teachers' ideas about scientific explanations, teachers' instructional scaffolding, particularly through discourse, and students' written explanations.

Students' Written Scientific Explanations

In our first research question, we asked, "How do third-grade students formulate written scientific explanations about seed structure and function?" Across the three classrooms, we found that a significant majority of student writing samples (66%) did not score for any facet of scientific explanations. Results from qualitative analysis of these student writing samples found that they were largely defined by data description without discussion of cause, effect, and mechanism. Writing samples that did illustrate some evidence of scientific explanations (34%) tended to include a cause, effect, and mechanism, though these samples had wide variations in the types of mechanisms attributed to the cause and effect.

Quantitative Analysis of Writing Samples. Across the three teachers and three lessons, there was a statistically significant difference among the presence of explanation-construction in the writing samples. We found a significant interaction between the teachers and lessons affecting the student written artifacts, $F [4, 62] = 10.23; p = .0367$. To discover the effect of the lessons and the teachers on students' written explanations, we used a one-way analysis of variance (ANOVA). We found a statistically significant difference across the teachers' enactments for Lesson 1, $F [2, 56] = 11.28, p = .0035$, and Lesson 3, $F [2, 56] = 9.17, p = .0102$. As shown in Figure 1, these differences are due in each instance to the high level of explanation-construction observed in students' written artifacts from Grace's classroom.

Next, to examine whether the interaction also occurred within each individual teacher's three lessons, we again used a one-way ANOVA to examine a single teacher across her three enactments. We found that each teacher was consistent across her three enactments for the presence of students written explanations (p values per teacher were $>.06$). Taken together, both sets of results indicate that, while lesson-specific opportunities for explanation-construction were similar across the three lessons for each teacher, opportunities afforded to students within each lesson varied significantly between the three teachers.

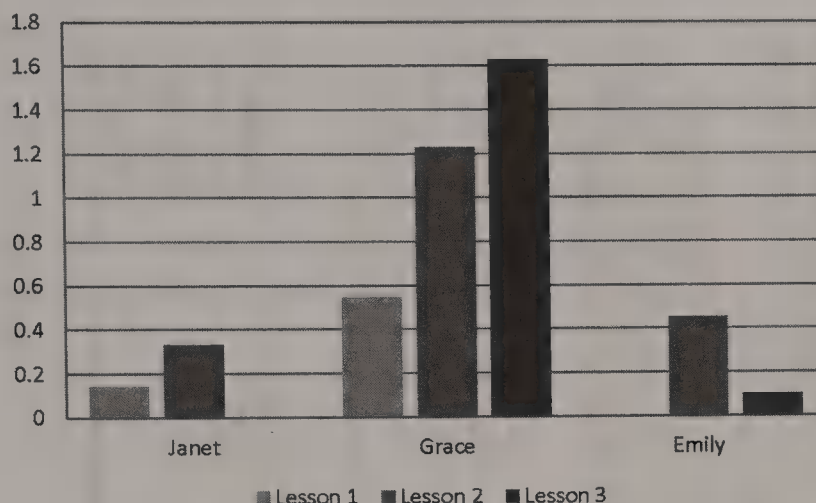


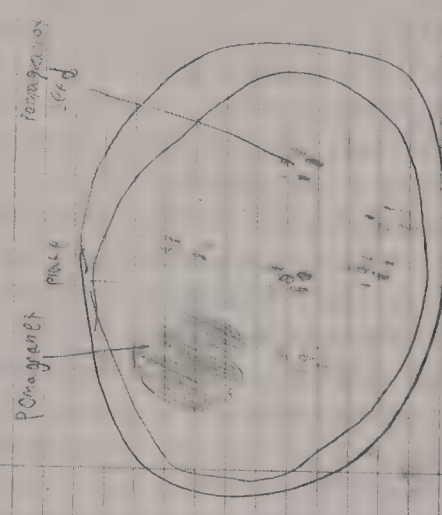
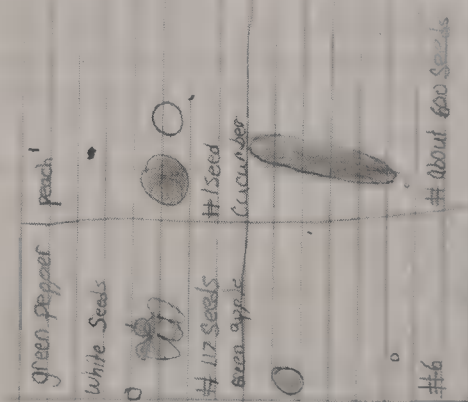




Figure 1. Mean scores for *formulating evidence-based explanations* in the student's writing samples across lessons and classrooms.

Qualitative Analysis of Writing Samples. As suggested in the results of quantitative analysis of student artifacts, our qualitative analysis find that that student writing artifacts from Grace's classroom most frequently exhibited one or more indicator of scientific explanations. The written explanations most frequently were *based on evidence* and *answer the investigation question*. Overall, we found few samples that coded for *propose new understanding* or *build on their existing knowledge*, but those few that contained these elements were also from Grace's classroom. We found few examples from either Janet's or Emily's classrooms that coded for any measure of explanation-construction.

Samples from Janet's classroom rarely coded for explanations because writing artifacts typically only contained recorded data collection from the lessons without an investigation question. All samples included a paragraph after data collection entries titled "After the Lesson." In these entries, students wrote freely but their writings rarely met the explanation measures. A typical "After the Lesson" response from Janet's classroom was a student reflection on the portion of the lesson they most enjoyed, which frequently did not include the investigation itself. For example, in Lesson 2 students were provided a range of fruits to observe and record properties, dissect, count (or estimate), and distinguish patterns among the different kinds of seeds. After the lesson, Janet provided opportunities for the students to taste all of the fruit. A common "After the Lesson" student entry for this lesson was a discussion of how the fruit tasted (Table 5). While eating the fruit was important as many students had never tasted the fruits examined during the lesson, it was not the key concept of this lesson, which was that different fruits have seeds of different size, shape and quantity.

Writing samples from Emily's classroom did include investigation questions, predictions (or hypothesis), data collection, and a conclusion (Table 5). Yet the written explanations from Emily's lessons rarely included mechanisms for *how* and *why* phenomena occurred. For example in Lesson 2 (Table 5), the artifacts from Emily's classroom recorded two investigation questions: "Where do seeds come from?" and "How are seed properties alike and different?" For example, the typical prediction statement this lesson was "I think some of the colors, texture, smell, size, and shape [of the seeds] will be different and alike" (S7, EL2) and a typical conclusion was "Properties of seeds are alike and different because some seeds can have the same size, color and shape." (S7, EL2). While these responses do *answer the investigation question*, they include no mechanism for *how* and *why* seeds come from fruit or seed properties are alike and different because the investigation questions they were posed did not require sense making to answer it.

TABLE 5
Student Artifact Samples for Lesson 2

	Grace (S7, GL2)	Emily (S2, EL2)	Janet (S10, JL2)
Question	How do the properties of the seeds compare in different kinds of fruits.	Where do seeds come from? How are seeds properties alike and different?	None
Prediction	They[compare different kinds of fruits because the bigger the fruit the bigger the seed. The smaller the fruit the smaller the seed	I think seeds come from fruits.	PREDICT:What will be inside. I think it would be a pea, flat spot's, smooth, hard, bumpy [bumpy], it could look bumpy [bumpy] but it could [be] smooth.
Observations			<div> <div> NAME OF FRUIT green bean NUMBER OF SEEDS six PROPERTIES OF SEEDS smooth light green 1.4# DRAWING OR SAMPLE OF SEEDS  </div> <div> NAME OF FRUIT peach NUMBER OF SEEDS 1 PROPERTIES OF SEEDS smooth light brown DRAWING OR SAMPLE OF SEEDS  </div> </div> <div> <div> NAME OF FRUIT peach NUMBER OF SEEDS 1 PROPERTIES OF SEEDS smooth light green DRAWING OR SAMPLE OF SEEDS  </div> <div> NAME OF FRUIT cucumber NUMBER OF SEEDS 3 PROPERTIES OF SEEDS smooth light brown DRAWING OR SAMPLE OF SEEDS  </div> </div>

(Continued)

The writing artifacts from Grace's classroom included investigation questions, predictions (or hypothesis), data collection and analysis, and a conclusion. The students' written conclusion sections most frequently included connections between cause, effect, and mechanism. For example, for Lesson 2, Grace's students recorded the investigation question, "How do the properties of the seeds compare in different kinds of fruits?" (S7, GL2). Their typical predictions discussed a specific property of the seed such as "The[y] [seed properties] compare . . . because the bigger the fruit the bigger the seed. The smaller the fruit the smaller the seed" (S7, GL2). Students then collected many observations of different kinds of fruits and drew picture graphs of the number of seeds found in each kind of fruit dissected (Table 5).

In their conclusion statements, the students typically exhibited sense making between the different fruits and different properties of seeds contained within each fruit. Our results indicate that two dominant themes occurred within the student artifacts from Grace's lessons. The first we considered as sophisticated explanations because the students made connections that were scientifically acceptable but outside of the range of content available to them during this lesson. For example, in Lesson 2 we found some students concluded that a correlation existed between the properties "seed size" and "seed number" (Table 5). The student appropriately connected the relationship between seed mass and seed number even though this relationship was not included in the lesson materials or enactment.

The second type of conclusion we found across all artifacts from all three of Grace's lessons we considered as naïve explanations because while students made connections between their evidence and explanation, their explanations were grounded in naïve mechanisms. For example, in Lesson 2, students' conclusions contained the notion that seed development in the plant was a constantly occurring process despite maturation and/or the fruit being removed from the plant. A typical response for this type of naïve explanation was "Some of the seeds were small. Some fruits had alot [sic] of seeds. Some fruits just had a c[o]uple . . . If there was only a c[o]uple small seeds it [the fruit] was not ful[l]y grown" (S3, GL2). While this conclusion statement is *based on evidence* and *answers the investigation question*, and the student shows evidence of sense making by providing an explanation for *why* some fruits had less seeds, the explanations is situated in a naïve mechanism that fruit maturation never stops occurring.

Elementary Teachers' Instructional Support of Explanation-Construction

In our second research question, we asked, "In what ways and why do third-grade teachers provide instructional support for students' formulation of scientific explanations about seed structure and function?" We analyzed the teachers' interviews and lesson enactments to explore how and why the teachers did (or did not) enact unit lessons to support their students to engage in scientific explanation-construction. Across the nine lessons, we identified 19 episodes of teacher-student discourse, mostly in Grace's classroom, where teachers enacted modified versions of unit lessons to support students to connect cause and effect with a mechanism. There were clear trends in the teachers' reasoning about explanation-construction that aligned with their instruction. Across the three teachers, our results illustrate two dominant models evident in the teachers reasoning and instructional practices around explanation-construction: correct explanations and self-as-explainer.

Emily and Janet: Correct Explanations. Emily and Janet viewed the curriculum as the locus of control. They both articulated that it was important for students to engage in "recipe-driven" (04:01:52) investigations because they need to be provided the scientific

facts. As Emily stated, sometimes a lesson has “got to be ‘here, let me tell you [what the answer should be]. Now you model it or draw it and then we will move forward’” (04:1:53) because, as she articulated, there was factual information students needed to know to engage in the investigations. Both teachers stressed the importance of students coming to the “correct” explanation that they interpreted as factual information. They also both emphasized they were concerned with time requirements to enact science lessons. They expressed concern that if they engaged their students with *why* questions and deviated from their curriculum materials, it would take a great deal of time for students to arrive at the scientific facts the lesson stressed.

We also found that both teachers struggled with what a scientific explanation might look like in the elementary classroom and whether it was something that should involve students determining whether their prediction was “right” or “wrong.” Janet and Emily emphasized correct explanations in both their verbal discussions with students and their support for students’ written explanations. Both focused their writing instructions on directing students to write what the investigation “should show” which they established from the curriculum materials (i.e., Lesson 1: seeds come from fruit, Lesson 2: different seeds have different properties, and Lesson 3: seeds sprout in the presence of water without soil).

As illustrated in the analysis of the student writing samples, Janet used her science notebooks the least of the three teachers. She was the only teacher that used the curriculum-provided data collection worksheets for all three lessons, and she read aloud from her curriculum materials prior to asking her students to write their “After the Lesson” statements. For example, in Lesson 2, after students had completed dissecting their fruits, she read directly from the curriculum materials and stated, “A plant part that contains seeds is called a fruit.” She then gave her students the writing prompt, “After the lesson I learned . . .,” and provided no further instructions for what students should consider or address in their writings. As observed in the student writing samples, this frequently resulted in students stating whether they were “correct” or “incorrect” in their conclusions and frequently did not score for explanation-construction (Figure 1; Table 5).

Unlike Janet, there were instances where Emily was able to provide instructional scaffolding to her students by affording them opportunities to formulate explanations. For example, in Lesson 3, students were examining the sprouting differences between four different types of seeds (bush bean, pea, and sunflower and popcorn seeds) in hydroponic growth. Their investigation question for this lesson was “What happens to seeds in water?” Emily supported a student in cause, effect, and the underlying unseen mechanism for why he was observing sprouts coming from the seed:

- T: What other changes have you noticed [to your lima bean]?
 S: It’s cracking open and grown a stem . . .
 T: What does this tell you about the seed?
 S: It’s changing a lot.
 T: Why is it changing a lot?
 S: Because of the water. The seeds are in water.
 T: OK. Why did the water change the seeds?
 S: Because the seeds need water to grow.
 T: Why?
 S: Um . . . because water . . . because seeds are living things. (Emily, 1a/7:53)

She prompted her student to connect the cause with the effect (“What other changes have you noticed?” and “Why is it changing a lot?”), then asked her student “Why?” two additional times leading this student to attribute a mechanism (“seeds are living things”) to the cause and effect which answered their investigation question of “what happens to seeds in water.”

However, this example of scaffolding a student to articulate a mechanism was a rare occurrence in Emily's classrooms and, as observed in the student artifacts, did not translate to students' writings (Figure 1). Like Janet, Emily struggled with what her students should write in the conclusion section of their science notebooks and how best to enact her curriculum when she asked students to write a conclusion. It was common in Emily's enacted lessons for her to ask students to use the conclusion section to "tell me whether or not your hypothesis was correct or incorrect" (Emily, 3c/3:43). However, while she identified the conclusion as the place for her students to determine whether their hypothesis was right or wrong, she also stated that the conclusion section was "something we need to work on" (04:5:26). She wondered if students should be writing more than whether their hypothesis was "right" or "wrong." When we asked her how she might work on the conclusion section with her students, she stated that she was not sure what the purpose of the conclusion section should be. As she stated, "I think I need to, myself, figure out exactly what I want" (04:7:24).

Grace: Self as Explainer. Grace defined good science teaching as not providing a single correct explanation. She expressed that science instruction should not focus on "just right or wrong" (01:1:92) answers and that her intention was for her students to experience a range of varying explanations. Furthermore, she stated that it was important for students to become "comfortable with the idea that it is very possible for there to be more than one explanation for why the seed numbers differ between the different kinds of fruit" (01:4:43) because, as she identified, there is not always a single correct answer in science. Grace discussed that if her students came to their own conclusions, then they had ownership of their reasoning and she had evidence of their learning.

She noted that she was working to integrate science with nonfiction writing so it was important to her that her students be provided sufficient time for writing in their science notebooks. Grace stated that during science notebook writing, students should consider how and why phenomena occurred because these are important components of nonfiction writing. However, Grace also identified that it was "tough" (01:1:24) to teach this way because it takes time and practice to get the students used to answering why questions and it also requires "thinking deep" (01:1:20) on the part of the teacher about the lesson content and the practices of science. Grace discussed that she takes her science lessons "deeper" through her goal for her students to "be responsible for their own learning" (01:1:10), which she identifies occurs through students writing their own conclusions. She stated that it was important that students write their own conclusions because it was "evidence for me of their learning" (01:2:66). Overall, we found that Grace viewed explanation-construction as a critical means to assess students' learning.

During her interviews, she expressed frustration with her science curriculum materials because she felt they did not meet her learning goals and did not support her students in understanding the "big picture" of the lesson. She discussed that if she was to do all of the parts of the curriculum, then the lessons would be time consuming and focused on vocabulary without sufficient opportunities for students to examine their data and evidence and write their conclusions. As a result, Grace modified each of her science lessons. For example, in Lesson 1, she chose to modify her lesson because she was concerned that the focus on the vocabulary term "properties" within the materials would detract from students' learning:

This lesson started out . . . with this apple and they were supposed to name the properties, so I was going to originally give them an apple. But, I decided not to. The only point of the

apple was to talk about properties. So, I scratched that part . . . we have to . . . get to the big picture. (01:02:24–30)

For this lesson, Grace said the “big picture” is the “amount of seeds in there [the pea pod] and that seeds are in fruit” (01:02:40). She expressed concern that any discussions that did not focus on these two concepts would distract student learning from the main idea. Instead, she decided to let her students begin the investigation of seed properties without an opening discussion.

Grace consistently supported her students to formulate explanations throughout her lessons but the mechanisms she accepted from students frequently did not match the targeted concept. Instead, we observed three different outcomes occurring in her teacher–student verbal explanation episodes. First, discussions in Grace’s classroom often resulted in students proposing elements of scientific explanations that demonstrated sophisticated reasoning about the phenomena, even though their proposed scientific explanations were based on concepts not covered in the lesson. For example, during Lesson 3, Grace’s students were observing the differences among their bush bean, pea, and sunflower and popcorn seeds placed in hydroponic growth in the week prior. Grace visited each individual student group to discuss the observations they were making about their seeds:

- T: What do you notice about your seeds? Are they sprouting the same way through?
- S: The beans are sprouting um they’re like twisted up together.
- T: Why do you think the beans are twisted up and the popcorn—
- S: Maybe it was like flipped over and it was like couldn’t grow down so maybe it got all twisted up when it was trying to grow down. (Grace, 3b/19:38)

In this student exchange, the student was concluding that the roots were attempting to grow toward the gravitational pull (i.e., gravitropism)—to “grow down”—even though the direction of root growth as well as the concept of gravity was not included in the lesson content. We considered this an example of sophisticated reasoning as the student was correct in their mechanism even though it was outside of the targeted lesson concept.

Second, we observed her verbally supporting student explanation-construction that included student evaluation of their prior understanding in light of new evidence. For example, in Lesson 3, she engaged an individual student in a discussion about seed growth who had verbally expressed at the beginning of the lesson that he thought a popcorn seed would only grow when planted in “hot stuff” (Grace, 3a/24:25) to grow a root. The student hypothesized that the popcorn seed would be unable to grow unless it was placed in the microwave to receive “hot stuff.” When Grace stopped at his table to question his current thinking about his hypothesis, the student identified that he no longer thought the popcorn seed needed to be in “hot stuff.” Grace continued the discussion prompting the student to provide her his evidence for his current thinking about how a popcorn seed grows. She also prompted this student for a mechanism for *how* and *why* the seed did not have to be in hot stuff in which he made connections that seeds are living things, and all seeds sprout in the presence of water and air.

Third, we observed 12 episodes across Grace’s three enacted lessons where students proposed naïve mechanisms grounded in the evidence from their investigation. This theme aligns with Grace’s students’ written explanations, which also included many naïve mechanisms. In each instance where these types of mechanisms occurred, we did not observe Grace attempt to question her students in identifying the scientifically accepted mechanism for the cause and effect they observed. For example, in Lesson 2, she asked students during a small group discussion why they thought seeds were located in different places inside different kinds of fruit. The students offered that “when they picked up the fruit, the juice

inside moved them [the seeds] around” (Grace, 2a/8:16). Grace offered that this was a “great explanation!” (Grace, 2a/8:16) and did not return to this explanation again over the course of the lesson enactments.

Summary

Overall, study findings suggest a link between the teachers’ ideas about scientific explanation construction during enactments, the presence of verbal explanation construction during their lessons, and student writing outcomes. Grace had the strongest conceptions of how scientific explanations may be enacted in the classroom, which she incorporated spontaneously in her lesson enactments. Moreover, her classroom had a statistically significant outcome for the increased presence of written explanations as compared to the other two classrooms. However, while Grace’s students’ explanations were more frequently grounded in evidence (i.e., scientific) and answered the investigation question, they also illustrated a range of naïve mechanisms related to seed structure and function.

SYNTHESIS AND DISCUSSION

Scientific explanation is a crucial scientific practice to foster elementary students’ conceptions about the world (Duschl et al., 2007; Metz, 2008; NRC, 2012). Prior studies examining elementary students’ engagement with scientific explanation-construction have shown some evidence of success. Yet it is important to note that these studies have predominantly relied upon researcher- and teacher-designed curriculum materials (e.g., “boutique” curriculum) to address these important facets of learning and instruction (e.g., Hapgood et al., 2004; Hardy et al., 2006; Herrenkohl et al., 1999; Mason, 2001; Metz, 2008). Off-the-shelf science curriculum materials may not foreground scientific reasoning (Metz, 2004), and researchers are only just beginning to examine the ways in which in-service teachers use these materials in elementary classrooms (Beyer & Davis, 2008; Biggers et al., 2013; Forbes & Davis, 2010; Forbes et al., 2013; Zangori et al., 2013). Further research is needed on the teacher–curriculum relationship so science educators might better understand ways to support both preservice and in-service teachers to plan with and enact these materials. The results presented here begin to shed light on how teachers use elementary science curriculum materials to foster students’ explanation-construction about plant growth and development. In particular, they suggest a link between teachers’ ideas about scientific explanations as a component of classroom inquiry, the instructional support they provide students for explanation-construction, and students’ written explanations for seed structure and function.

First, students’ written explanations show that while a small number of students were engaging in scientific reasoning, we found few scientific explanations in students’ written work. The majority of students’ written samples were composed of observations or statements without evidence. These types of statements do not only constitute scientific understanding because the students have not been asked to engage in knowledge production, only to provide a written account of their observations (Duschl et al., 2007; Braaten & Windschitl, 2011; Metz, 2008; Salmon, 1998). While the students here had multiple and lengthy opportunities to engage in hands-on activities with seeds, fruit, and seed growth, their written work did not indicate that they engaged in sense making about their investigations (Biggers et al., 2013; Forbes et al., 2013; NRC, 2000, 2012; Zangori et al., 2013). This finding suggests that despite calls for providing student opportunities to develop conceptual understanding of plant growth and development through multiple and varied hands-on engagement opportunities with plants (Barman et al., 2006; Canal, 1999; Jewel, 2002; Patrick

& Tunnicliffe, 2004), such opportunities were relatively limited and must be carefully crafted to promote students' learning.

The small portion of students that were afforded opportunities to engage in scientific reasoning did so in varied ways. These instances when cause, effect, and mechanism were generated by the third graders were grounded in evidence, answered an empirical question, and in some instances, built upon and extended students' understanding. Each of these are critical epistemic commitments of explanation-construction articulated by the NRC (2000, 2012) and emphasized in past research (Biggers et al., 2013; Forbes et al., 2013; Mason, 2001; Ryu & Sandoval, 2012; Zangori et al., 2013). Since early learners' reasoning is constrained by the domain-specific conceptual knowledge they have available, in its absence, students will rely upon intuitive patterns in their evidence to generate explanations (Duschl et al., 2007; Schauble, 1996; Zuzovsky & Tamir, 1999). However, plant structure and function is a complex biological system made up of many underlying, unobservable mechanisms and connections between abiotic and biotic factors (Patrick & Tunnicliffe, 2011; Schussler, 2008; Wandersee & Schussler, 1999). While this was occasionally fruitful, such as when students recognized the correlation between seed size and seed number, these instances were rare. More frequently, if students engaged in scientific reasoning, they attributed alternate conceptions to their evidence about seeds, fruit, and seed growth, as has been found in prior research about students' alternate conceptions about plant growth and development (Barman et al., 2006; Canal, 1999; Jewel, 2002).

Second, this study sheds light on the teachers' conceptions and orientations toward explanation-construction as a scientific practice. On the one hand, when teachers identify with the "correct explanation" model (Horwood, 1988; Zuzovsky & Tamir, 1999), they may not provide support or opportunities for their students to engage in sense making. This may be because the teachers assume that they have provided their students with the "right" information from the curriculum so no further engagement in scientific explanations or evaluation of scientific explanations is necessary. However, in this manner, the science content becomes separated from knowledge-building processes (Braaten & Windschilt, 2011; Metz, 2008).

On the other hand, when teachers identify as "self-as-explainer" (Horwood, 1988; Zuzovsky & Tamir, 1999), the reasoning activity itself, without concern of content, becomes the goal of the lesson. The focus within this model is on student generation of *any* mechanism for cause and effect regardless of its scientific acceptability. While this model may engage students in scientific reasoning, it does not include consideration of the manner in which students interpreted their data and evidence for cause and effect or the scientific acceptability of *how* and *why* the phenomenon occurred. Teachers who align with this model view the epistemic practice of science as separate from domain-specific knowledge (Metz, 2008; Ryu & Sandoval, 2012).

Yet, despite these different ideas about scientific explanations, our results indicate commonality among the three teachers' views as to *where* scientific explanations fit within instruction. While the teachers had different ideas about the ways scientific explanations should be incorporated into their enacted lessons, all three teachers were uniform in their view that the scientific explanation is the outcome and conclusion to the investigation. In other words, the teachers viewed scientific explanation is the means to an end for the science lesson (Beyer & Davis, 2008). This is problematic as it does not provide students opportunities to compare and evaluate their explanations, an uncommon practice in elementary science classrooms (Biggers et al., 2013). Without this crucial next step, student opportunities to engage in knowledge production and develop scientific understanding are severely limited (Duschl et al., 2007; Jewel, 2002; Mason, 2001; NRC, 2000, 2012; Schauble, 1996).

Finally, third, study findings suggest that both the instructional support and scaffolding teachers provided students to formulate explanations reflect their ideas and conceptions of scientific explanations in elementary science learning environments. All three teachers had to depend on their own ideas about scientific explanations and student learning because their materials did not support them in scaffolding this practice within their lessons. Prior research suggests that elementary students can effectively engage in scientific reasoning when provided synergistic support (Tabak, 2004) from the teacher, their curriculum materials, and through classroom discourse (Hardy et al., 2006; Herrenkohl et al., 1999; Mason, 2001; Metz, 2008, 2009; Ryu & Sandoval, 2012; Zuzovsky & Tamir, 1999). However, a core component of inquiry investigations is providing students opportunities to generate explanations and compare and evaluate their explanations against prior knowledge, other students' explanations, and/or other scientific explanations (NRC, 2000, 2012). In this manner, students have opportunities to engage in the practices of science and examine their preexisting mechanism libraries for adequacy in explaining the phenomenon (Duschl et al., 2007; Mason, 2001; Schauble, 1996).

This approach to explanation construction is the "structure of science" explanation model (Horwood, 1988; Zuzovsky & Tamir, 1999), which we did not observe with the teachers studied here. This model of explanation-construction provides a middle point along the more teacher- or student-directed continuum of explanation construction. This may be a more advantageous model for both teachers and students to engage in explanation-construction because it provides opportunities for students to generate explanations, but then supports students to evaluate those explanations until a single satisfactory explanation is constructed. As the NRC 2012 study suggests, "... knowing why the wrong answer is wrong can help secure a deeper and stronger understanding of why the right answer is right" (p. 44) and the "structure-of-science" model would support teachers in helping their students develop scientific understanding in each instance. However, this requires that teachers use their curriculum materials in a flexible manner modifying the materials where they, and their students, require support in connecting cause, effect, and the scientifically acceptable mechanism as a learning goal of the lesson.

IMPLICATIONS

Addressing all components of scientific explanations is required for conceptual change (NRC, 2012), but not all facets of explanation-construction are intuitive (Braaten & Windschitl, 2011; Duschl et al., 2007). While elementary science curriculum materials should afford students meaningful sensemaking opportunities through engagement with plant phenomena in ways that link students' prior knowledge to targeted concepts and challenge student's alternate conceptions, these opportunities are atypical (Biggers et al., 2013; Forbes et al., 2013; Metz, 2004, 2008; Ryu & Sandoval, 2012; Zangori et al., 2013). To create a coherent learning experience that aligns with the practices of science, elementary science learning environments must be composed of scaffolds that work synergistically (Tabak, 2004) with the curriculum, associated tools such as science notebooks, and instruction. For this to occur, an emphasis on process and content must be carefully intertwined (Herrenkohl et al., 1999; Metz, 2008; Schauble, 1996). Therefore, these findings have important implications for both professional development and curriculum development.

Curriculum materials should be designed to afford students opportunities to formulate scientific explanations that connect cause, effect, and mechanism (NRC, 2012). Although no formal review of elementary science curriculum materials has been conducted (Kesidou & Roseman, 2002), results of some studies suggest that elementary science curriculum materials underemphasize explanation-construction (Biggers et al., 2013; Forbes et al.,

2013; Zangori et al., 2013). In these instances, the materials focus heavily on engagement with phenomena and data collection but typically do not ask students to consider causal inferences. For elementary students to be successful in developing scientific understanding of plant structure and function, students should have opportunities to focus on the key processes with the underlying unobservable mechanisms made explicit. For example, the curriculum studied here provided surface explanations such as “seeds are contained in fruit,” without providing students access to the underlying causal mechanism for *how* and *why* seeds are contained in fruit. Without providing students access to this fundamental function of seeds within the plant life cycle, then the process appears “magical” because students do not have a reason for why they are finding seeds in fruit (Schussler, 2008). Furthermore, students require opportunities to compare and evaluate their explanations with other scientific explanations to arrive at consensus of the acceptable scientific mechanism (Ryu & Sandoval, 2012). Without providing students access to these critical facets of explanation-construction, then their preexisting mechanism libraries about seed structure and function will not be challenged and naïve conceptions about plant growth and development will remain.

Effective curriculum materials should also include components that are educative for teachers themselves (Davis & Krajcik, 2005). If curriculum is educative for the ways in which to support students in practices of science and the conceptions included within the materials, teachers will be better equipped to understand where student conceptual knowledge may need to be bolstered and make lesson modifications as needed. Curriculum materials should also provide teachers with ideas about the typical alternate conceptions students bring with them about plant growth and development, so they are prepared to engage students in examining these naïve mechanisms and develop new understanding (Barman et al., 2006; Canal, 1999; Jewel, 2002; Patrick & Tunnicliffe, 2004). Educative curriculum materials would provide examples of early learners’ engagement with scientific reasoning and what these episodes can look like in the elementary classroom. Furthermore, the nature of educative materials is to support teachers’ pedagogical reasoning about the tasks included in the lesson rather than for teachers to use the materials as a script (Davis & Krajcik, 2005). In this manner, teachers can be flexible with their materials (Remillard, 2005) and determine where opportunities for sense making would be most appropriate and the ways in which to engage their students in these opportunities.

CONCLUSIONS

This study contributes to a limited body of research on elementary students’ learning about plants and explanation-construction in elementary science learning environment, which provides possible explanations as to why students’ naïve mechanisms about plant structure and function persist after almost three decades of research (e.g., Barman et al., 2006; Canal, 1999; Jewel, 2002; Patrick & Tunnicliffe, 2011; Wandersee & Schussler, 1999). Elementary students’ opportunities for active sense making (Mason, 2001; Ryu & Sandoval, 2012) are often rare, and, when included, may not result in students replacing naïve mechanisms with scientifically acceptable ones (Hardy et al., 2006). Despite calls for providing students with opportunities to develop conceptual understanding of plant growth and development through investigation (Barman et al., 2006; Canal, 1999), findings from this study suggest that such experiences in the elementary grades may not always foster and promote conceptual change.

This work supports and extends the notion that elementary teachers’ pedagogical understanding of explanation-construction and science content requires strengthening, which has implications for science educators and curriculum developers. Future research should

explore other widely available off-the-shelf elementary curriculum materials for inclusion of scientific-explanation construction and the coherence of science content to explore whether this issue is endemic across elementary curriculum materials. Future work should also investigate how widely used curriculum materials may be supplemented to support both teacher and student learning in formulating scientific explanations and meeting the new science education standards (NRC, 2012) that include both modeling and argumentation frameworks—neither of which was present in the curriculum studied here. Such work will help science teacher educators and curriculum developers better understand how to support teachers to promote explanation construction in their classrooms.

This research is funded by the Roy J. Carver Charitable Trust. However, any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors. We appreciate the interest and cooperation of Kim Gasaway and the elementary teachers who made this research possible. We also thank Mandy Biggers, Sheila Barron, Lindsey Zahn, Madison Fontana, and anonymous reviewers for their help in thinking about these issues and their thoughtful comments on earlier versions of this paper.

REFERENCES

- Barman, C. R., Stein, M., McNair, S., & Barman, N. S. (2006). Students' ideas about plants and plant growth. *The American Biology Teacher*, 68(2), 73–79.
- Beyer, C. J., & Davis, E. A. (2008). Fostering second graders' scientific explanations: A beginning elementary teacher's knowledge, beliefs, and practice. *Journal of the Learning Sciences*, 17(3), 381–414.
- Beyer, C. J., Delgado, C., Davis, E. A., & Krajcik, J. (2009). Investigating teacher learning supports in high school biology curricular materials. *Journal of Research in Science Teaching*, 46(9), 977–998.
- Biggers, M., Forbes, C. T., & Zangori, L. (2013). Elementary teachers' curriculum design and pedagogical reasoning for supporting students' comparison and evaluation of evidence-based explanations. *The Elementary School Journal*, 114(1), 48–72.
- Braaten, M., & Windschitl, M. (2011). Working toward a stronger conceptualization of scientific explanation for science education. *Science Education*, 95(4), 639–669.
- Canal, P. (1999). Photosynthesis and “inverse respiration” in plants: An inevitable misconception? *International Journal of Science Education*, 21(4), 363–371.
- Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.
- Davis, E. A., & Krajcik, J. (2005). Designing educative curriculum materials to promote teacher learning. *Educational Researcher*, 34(3), 3–14.
- Denzin, N. K., & Lincoln, Y. S. (2000). The discipline and practice of qualitative research. In N. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 1–28). Thousand Oaks, CA: Sage.
- Duschl, R. A., Schweingruber, H. A., & Schouse, A. W. (Eds.). (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: National Academy Press.
- Enyedy, N., & Goldberg, J. (2004). Inquiry in interaction: How local adaptations of curricula shape classroom communities. *Journal of Research in Science Teaching*, 41(9), 905–935.
- Forbes, C. T., Biggers, M., & Zangori, L. (2013). Investigating essential characteristics of scientific practices in elementary science learning environments: The Practices of Science Observation Protocol (P-SOP). *School Science and Mathematics*, 113(4), 180–190.
- Forbes, C. T., & Davis, E. A. (2010). Beginning elementary teachers' beliefs about the use of anchoring questions in science: A longitudinal study. *Science Education*, 94(2), 365–387.
- FOSS. (2005). *Teacher guide: Structures of life*. Berkeley, CA: Delta Education.
- Hapgood, S., Magnusson, S. J., & Palincsar, A. S. (2004). Teacher, text, and experience: A case of young children's scientific inquiry. *Journal of the Learning Sciences*, 13(4), 455–505.
- Hardy, I., Jonen, A., Möller, K., & Stern, E. (2006). Effects of instructional support within constructivist learning environments for elementary school students' understanding of “floating and sinking.” *Journal of Educational Psychology*, 98(2), 307–325.
- Herrenkohl, L. R., Palincsar, A. M., DeWater, L. S., & Kawasaki, K. (1999). Developing scientific communities in classrooms: A sociocognitive approach. *Journal of the Learning Sciences*, 8(3-4), 451–493.

- Horwood, R. H. (1988). Explanation and description in science teaching. *Science Education*, 72(1), 41–49.
- Jewel, N. (2002). Examining children's models of seed. *Journal of Biological Education*, 36(3), 116–122.
- Kesidou, S., & Rosemann, J. E. (2002). Do middle school science programs measure up? Findings from Project 2061's curriculum review. *Journal of Research in Science Teaching*, 39(6), 522–549.
- Kleinbaum, D. G., Kupper, L. L., Muller, K. E., & Nizam, A. (1998). *Applied regression analysis and other multivariable methods*. New York: Duxbury Press.
- Kuhn, D. (2009). Do students need to be taught how to reason? *Educational Research Review*, 4(1), 1–6.
- Lowery L. (1998). *The biological basis of thinking and learning* [monograph]. Full Option Science System. Berkeley, CA: Lawrence Hall of Science. Retrieved June 20, 2012, from http://lhsfoss.org/newsletters/archive/pdfs/FOSS_BBTL.pdf.
- Mason, L. (2001). Introducing talk and writing for conceptual change: A classroom study. *Learning and Instruction*, 11(4–5), 305–329.
- Metz, K. E. (2004). The knowledge building enterprises in science and elementary school science classrooms. In L. B. Flick & N. G. Lederman (Eds.), *Scientific inquiry and nature of science* (pp. 105–130). Dordrecht, The Netherlands: Springer.
- Metz, K. (2008). Narrowing the gulf between the practices of science and the elementary school science classroom. *The Elementary School Journal*, 109(2), 138–161.
- Metz, K. (2009). Elementary school teachers as “targets and agents of change”: Teachers' learning interaction with reform science curriculum. *Science Education*, 93(5), 915–954.
- National Research Council. (2000). *Inquiry and the National Science Education Standards: A guide for teaching and learning*. Washington, DC: The National Academy Press.
- National Research Council. (2012). *A framework for K-12 science education: Practices, cross-cutting concepts, and core ideas*. Washington, DC: National Academies Press.
- Patrick, P., & Tunnicliffe, S. D. (2011). What plants and animals do early childhood and primary students' name? Where do they see them? *Journal of Science Education and Technology*, 20(5), 630–642.
- Patton, M. Q. (2001). *Qualitative research and evaluation methods*. Thousand Oaks, CA: Sage.
- Remillard, J. T. (2005). Examining key concepts in research on teachers' use of mathematics curricula. *Review of Educational Research*, 75(2), 211–246.
- Ryan, G. W., & Bernard, H. R. (2000). Data management and analysis methods. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp 769–802). Thousand Oaks, CA: Sage.
- Ryu, S., & Sandoval, W. A. (2012). Improvements to elementary children's epistemic understanding from sustained argumentation. *Science Education*, 96(3), 488–526.
- Salmon, W. (1998). *Causality and explanation*. New York: Oxford University Press.
- Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology*, 32(1), 102–119.
- Schussler, E. (2008). From flowers to fruits: How children's books represent plant reproduction. *International Journal of Science Education*, 30(12), 1677–1696.
- Statistical Analysis System (2013). [Computer software.] Cary, NC: SAS Institute.
- Stern, L., & Rosemann, J. E. (2004). Can middle-school science textbooks help students learn important ideas? Findings from project 2061's curriculum evaluation study: Life science. *Journal of Research in Science Teaching*, 41(6), 538–568.
- Tabak, I. (2004). Synergy: A complement to emerging patterns of distributed scaffolding. *Journal of the Learning Sciences*, 13(3), 305–335.
- Wandersee, J., & Schussler, E. E. (1999). Preventing plant blindness. *The American Biology Teacher*, 61(2), 82–86.
- Zangori, L., Forbes, C., & Biggers, M. (2012). This is inquiry . . . right? *Science and Children*, 50(1), 48–53.
- Zangori, L., Forbes, C. T., & Biggers, M. (2013). Fostering student sense making in elementary science learning environments: Elementary teachers' use of science curriculum materials to promote explanation construction. *Journal of Research in Science Teaching*, 50(8), 989–1017.
- Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, 27(2), 172–223.
- Zuzovsky, R., & Tamir, P. (1999). Growth patterns in students' ability to supply scientific explanations: Findings from the Third International Mathematics and Science Study in Israel. *International Journal of Science Education*, 21(10), 1101–1121.

Investigating the Link Between Learning Progressions and Classroom Assessment

ERIN MARIE FURTAK,¹ DEB MORRISON,¹ HEIDI KROOG²

¹*School of Education, University of Colorado at Boulder, Boulder, CO 80309, USA;*

²*School of Education and Human Development, University of Colorado Denver, Denver, CO 80217, USA*

Received 7 May 2013; accepted 3 April 2014

DOI 10.1002/sce.21122

Published online 18 June 2014 in Wiley Online Library (wileyonlinelibrary.com).

ABSTRACT: An increasing number of researchers are calling for learning progressions to be used as interpretive frameworks for teachers conducting classroom assessment. The argument posits that by linking classroom assessments to learning progressions, teachers will have better resources to interpret and take instructional action on the basis of what students know. In this paper, we draw on data from a research project in which we have supported high school biology teachers in interpreting student responses to a multiple-choice classroom assessment linked to a learning progression for natural selection. We draw upon multiple sources of data, including student responses to a pre–post classroom assessment, artifacts from professional development sessions, and videotapes and transcripts of professional development meetings to construct a case study of one department of teachers as they came to understand the learning progression, interpreted results of classroom assessments, and revised their instruction. We use this case to illustrate the promise and challenges associated with linking classroom assessments to learning progressions. We conclude with recommendations to the field and suggestions for future work in this area. © 2014 Wiley Periodicals, Inc. *Sci Ed* 98:640–673, 2014

Correspondence to: Erin Marie Furtak; e-mail: erin.furtak@colorado.edu

The findings in this paper were originally presented at the National Association of Research on Science Teaching Annual International Conference, Puerto Rico, April 2013.

Contract grant sponsor: National Science Foundation.

Contract grant number: 0953375.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

INTRODUCTION

Research on learning progressions has influenced much of the work in science education over the past decade, including standards initiatives, curriculum development, and assessment design (Cocoran, Mosher, & Rogat, 2009). In addition to being used for these purposes, some researchers have also argued that as representations of how student understanding develops in a domain, learning progressions may be useful tools for teachers to support their classroom assessment practices. Such cognitive domain models would help teachers understand how core ideas in science are related, as well as the common everyday ideas students bring to the classroom, and better prepare them to elicit and respond to student ideas during instruction (Bennett, 2011; Furtak, 2012; Heritage, 2008). While many learning progressions have been developed in recent years (see Duschl, Maeng, & Sezen, 2011 for a summary), the ways in which they might support teachers in conducting and interpreting classroom assessments have not yet been fully explored.

This paper chronicles our investigation of the ways in which a group of high school science teachers interpreted classroom assessments linked to a learning progression on natural selection. We begin this report with a review of research on learning progressions and classroom assessment and discuss ways in which the two may work together to support teachers' interpretations of classroom assessment results. Then, we describe the natural selection learning progression and associated assessments that we developed. We draw upon multiple sources of data collected at one school site to describe both the potential utility and the challenges we have encountered in direct use of a learning progression in the classroom. We conclude by identifying implications for the future use of learning progressions to support teachers' interpretation of assessment results.

LEARNING PROGRESSIONS AND CLASSROOM ASSESSMENT

The science education research community has recently devoted considerable effort to develop learning progressions, or what Corcoran, Mosher, and Rogat (2009) defined as "hypothesized descriptions of the successively more sophisticated ways student thinking about an important domain of knowledge or practice develops as children learn about and investigate that domain over an appropriate span of time." Learning progressions represent some kind of developmental sequence that emphasizes the connectedness of concepts, practices, or the interrelationship of both within a domain (Songer, Kelcey, & Gotwals, 2009).

Several research groups have illustrated the symbiotic nature of learning progressions and assessment development, where analyses of student work and interviews can lead to draft learning progressions, which are used as foundations for assessment design, and then the results of the assessment in turn help to inform and refine the learning progression. For example, Smith, Wiser, Anderson, and Krajcik (2006) developed a learning progression for students' understanding of matter and atomic molecular theory and used it to develop sample items. Many of the assessments linked to learning progressions involve open-ended items and interviews (e.g., Mohan, Chen, & Anderson, 2009; Duncan, Rogat, & Yarden, 2009; Jin, Choi, & Anderson, 2009), although some involve multiple-choice items (e.g., Songer et al. 2009). Several of the preceding authors acknowledged the complexity in assembling assessments linked to learning progressions, as well as the difficulty in analyzing the data from such assessments relative to a learning progression.

Furtak (2009, 2012) has argued that, as representations of how student thinking develops in a domain, learning progressions may also support teachers in identifying and responding to student thinking during classroom instruction. In fact, when accompanied by

assessments linked to different developmental levels, learning progressions have the potential to support teachers in interpreting assessment results and suggesting instructional paths forward (Mohan & Anderson, 2009). According to Corcoran et al. (2009), assessments based on learning progressions could “provide information that is more easily interpreted by teachers and potentially allow them to make better informed and more precise decisions about student needs and how to respond to them instructionally” (p. 23).

Assessment items can be specifically designed with links to different levels in a learning progression or with varying levels of scaffolding so that patterns of student responses can provide information about what students know and can do relative to the learning progression (e.g., Briggs et al., 2008; Sadler, 1998; Songer et al., 2009; Wilson & Sloane, 2000). Furthermore, if student responses are aligned to levels on a learning progression, teachers can easily see what students currently know, what they have learned, and where they might go next. Learning progressions that combine information about how ideas or practices develop in a domain with sequences that are linked with the everyday ideas students have prior to instruction may in particular help teachers interpret assessment results (Furtak, 2012).

Based on these potential advantages, a growing number of researchers have called for the use of learning progressions to support teachers’ classroom assessment practices. Bennett (2011) argued that models for how student thinking develops in a domain might help teachers identify and make inferences about what students know. Similarly, Heritage (2008) suggested that having maps of how student understanding develops in a domain could help teachers better provide feedback to learners to move them forward.

Mark Wilson and the Berkeley Education Assessment Research Group (BEAR) have performed the most prominent work in the area of linking representations of the development of student ideas and assessments. As Wilson and Sloane (2000) described, the BEAR assessment system provides teachers a set of tools linked back to a map of how particular student ideas develop through a course of instruction, what Wilson has elsewhere referred to as a construct map. Wilson (2005) defined a construct map as “a coherent and substantive definition for the content of the construct, and an idea that the construct is composed of an underlying continuum” (Wilson, 2005, p. 26). The BEAR approach to embedded classroom assessment, also called the BEAR Assessment System, has been used to develop construct maps in a number of disciplines, including the carbon cycle, chemistry, rational number reasoning, and statistical reasoning (see Draney, 2009, e.g.).

Minstrell’s work with the DIAGNOSER software program is another prominent effort to link student responses to representations of student ideas in a domain through classroom assessment (Minstrell, Anderson, Kraus, & Minstrell, 2008). In this program, students respond to multiple choice questions with distractors that are matched to different types of student ideas within a domain, which Minstrell calls facets. Facets are not organized in the same style as is found in learning progressions, but are categorized in clusters roughly grouped according to ideas that are more or less sophisticated. Gold Facets, as Minstrell calls them, are specific learning goal statements, and problematic facets describe naïve and everyday ideas. Teachers can look at student responses to DIAGNOSER questions and identify which of the facets best describe what students currently know. Thus, DIAGNOSER assessments are designed to document the range of student ideas in a particular domain in a manner that is useful for instruction.

Summary of Previous Work

Although Wilson’s (2009) and Minstrell et al.’s (2008) work have illustrated the ways in which teachers might use representations of student understanding to interpret classroom

assessment results, the field is still exploring how such representations actually support teachers in their daily instruction. Corcoran et al. (2009) noted that learning progressions are “as yet unproven tools for improving teaching and learning, and . . . developing and utilizing this potential poses some challenges” (p. 5). Many questions could be raised about the nature of the support learning progressions might provide to teachers, and the challenges that might follow. For example, researchers studying learning progressions frequently argue that they represent multiple trajectories or pathways for learning (National Research Council [NRC], 2007); as such, how are teachers to interpret assessment information linked to them to guide their classroom practice? Corcoran et al. (2009) also argued that although learning progressions suggest simple content domains, those content domains are necessarily interwoven with other conceptual understandings. The question then follows, how can we support teachers to understand the nature of learning progressions created in complex, interconnected conceptual domains?

Finally, questions remain as to the value of learning progressions in scaffolding teachers’ perceptions of student ideas in science. Teachers commonly view student ideas in science as correct or incorrect (Otero & Nathan, 2008) or attend to superficial elements of student reasoning rather than the nature and substance of student thinking (e.g., Coffey, Hammer, Levin, & Grant, 2011). Bennett (2011) argued that to effectively conduct classroom assessment, teachers should learn to make “distinctions among [the] errors, slips, misconceptions, and lack of understanding” (p. 17) that underlie a student’s response. Furthermore, Heritage, Kim, Vendlinski, and Herman (2009) found that mathematics teachers were able to more effectively estimate the level of student ideas than to plan subsequent classroom instruction. Both Bennett (2011) and Heritage et al. (2009) noted that learning progressions may help to scaffold teachers’ understandings of student ideas. However, Furtak (2012) found that teachers may interpret learning progressions as reinforcing “right–wrong” perceptions of student ideas.

We have taken up these and other issues over the past 3 years in which we have studied teachers’ use of a learning progression as a tool to interpret student responses to a classroom assessment. Several authors have called for such use of learning progressions (e.g., Bennett, 2011; Heritage, 2008; Heritage et al., 2009), and in this paper we examine empirical evidence of how teachers engage with learning progressions to support their classroom assessment practices. We draw on multiple sources of data from our project to discuss the potential utility of the learning progression–classroom assessment link, as well as challenges we have faced as we introduced teachers to learning progressions, and identify ways in which future work might respond to these challenges.

STUDY CONTEXT

This paper reports on results from an ongoing National Science Foundation–funded study called Elevate: Educative Learning Progressions for Teacher Development. The Elevate study focused on the scientific concept of natural selection, one of the disciplinary core ideas in biology (Board on Science Education [BOSE], 2012; Dobzhansky, 1973). We have developed a multidimensional learning progression for this concept, a classroom assessment to which it is linked, and have studied the use of these tools by a group of high school biology teachers over the course of a multiyear on-site investigation. In this section, we will describe the learning progression, the professional development program in which we worked to support teachers in using these tools as part of their classroom instruction, the participants in the study, and the Daphne Assessment of Natural Selection (DANS), the instrument to which the learning progression is linked.

Prior Representations of Student Ideas About Natural Selection

Studies in psychology and education have documented that individuals of all ages have difficulties understanding the concept of evolution by natural selection (Anderson, Fisher, & Norman, 2002; Rudolph & Stewart, 1998). For example, students frequently have trouble recognizing the role of random molecular processes in science (Ferrari & Chi, 1998; Odom & Barrow, 1995), are likely to believe that new traits arise as a result of an organism's needs rather than occurring through random genetic processes (Bishop & Anderson, 1990; Dagher & Boujaoude, 2005; Geraedts & Boersma, 2006), and confuse everyday meanings of "fitness" with its biological definition (Anderson et al., 2002). Teaching natural selection is further complicated by the fact that students and adults have been shown to exhibit patterns of mixed reasoning in which different kinds of naïve ideas are mixed with scientifically accurate ideas in explanations (Evans et al. 2010; Nehm & Ha, 2011).

A number of researchers have articulated learning progressions for natural selection and to identify the subdimensions of this concept (Evans et al. 2010; Lehrer & Schauble, 2012; Metz, Sisk-Hilton, Berson, & Ly, 2010). For example, Catley, Lehrer, and Reiser (2005) set out to identify "How learning performances oriented around foundational concepts (big ideas) of evolution can articulate temporal sequences supporting students' long-term cognitive development" (p. 6). They parsed natural selection into six conceptual structures (diversity, structure–function, ecology/interrelationships, variation, change, and geologic processes) and two additional core conceptual structures, which they termed habits of mind (forms of argument and mathematical tools). They then identified potential learning progressions and provided rationales for the way they represented the big ideas for each of the conceptual structures and habits of mind within grade bands K–2, 3–5, and 6–8. Each grade band also included suggested learning performances with sample activities.

In a study of how elementary school students model "big ideas" that can serve as conceptual foundations for their later understanding of evolution, Lehrer and Schauble (2012) articulated a learning progression in which pathways are identified for how students learn to model three strands: variability, change, and ecosystems in Grades K through 6. Lehrer and Schauble (2012) then created construct maps in each of four categories (change was disaggregated into individual and population change) and identified "benchmarks that constitute especially consequential shifts in student thinking" (p. 713). Each construct map included five to nine levels and provided a level of detail that was suitable to serve as a guide for instruction. Each benchmark was illustrated with learning performances at that level and accompanied by examples. At the same time, Metz et al. (2010) delineated a learning progression for students' understanding of the conceptual underpinnings of evolution in Grades 2–3. This learning progression included seven levels that began with where they expected students to enter school (that organisms live where they belong and that they live where they get what they need) and listed increasingly powerful and complex explanations up to a top level encompassing natural selection's outcome (organisms are well adapted to where they live). Metz et al. (2010) also developed a secondary progression for life cycle, resemblance of parent and offspring, and inheritance.

Other related work has helped to identify dimensions of natural selection. For example, Shulman (2006) explored how certain student ideas parallel pre-Darwin theories about the origin of species, defining these "transformationist" misconceptions as attributing change to "a single process operating on a species' 'essence'" (p. 173) and juxtaposed them with more sophisticated "variational" reasoning. He articulated how these two types of ideas manifest themselves within six evolutionary phenomena, including variation, inheritance, adaptation, domestication, speciation, and extinction. In a study of museum visitors, Evans et al. (2010) identified patterns of mixed evolutionary reasoning among museum visitors,

including informed naturalistic reasoning, which included spontaneous mention of the key evolutionary concepts of variation, inheritance, selection, and time: novice naturalistic reasoning, which include intuitive or everyday explanations that are also given by young children, and creationist reasoning, or supernatural explanations.

The Elevate Learning Progression

Similar to Catley et al. (2005) and Lehrer and Schauble (2012), we have articulated multiple dimensions for the concept of natural selection, building on what we view as core elements of Catley et al.'s (2005) progression (ecology/interrelationships, variation, and change). On the basis of input provided by practicing biology teachers in a previous study (see Furtak, 2009), we adopted Mayr's (1982) sequence of facts and inferences as a disaggregation of the multiple dimensions of natural selection. Mayr's account states in simple terms the understandings a high school student may assemble to provide a well-articulated explanation of natural selection. Mayr's "Five Facts and Three Inferences" form the core of the *National Science Education Standards* (NRC, 1996) description of evolution and are as follows:

Fact 1: All species have such great potential fertility that their population size would increase exponentially if all individuals that are born would again reproduce successfully.

Fact 2: Except for minor annual fluctuations and occasional major fluctuations, populations normally display stability.

Fact 3: Natural resources are limited. In a stable environment they remain relatively constant.

Inference 1: Not all offspring survive to reproductive age in part because of competition for natural resources.

Fact 4: No two individuals are exactly the same; rather, every population displays enormous variability.

Fact 5: Much of this variation is heritable.

Inference 2: Survival in the struggle for existence is not random but depends in part on the hereditary constitution of the surviving individuals. This unequal survival constitutes a process of natural selection.

Inference 3: Over the generations this process of natural selection will lead to a continuing gradual change of populations, that is, to evolution and to the production of new species (Mayr, 1982, p. 479–480).

Mayr's account states in simple terms the stepwise understandings one needs to have, and the inferences one needs to make on the basis of those understandings, to understand the whole process of natural selection. These "Five Facts and Three Inferences" are similar to the disciplinary core ideas in the *Next Generation Science Standards* (NGSS; LS4.B: Natural Selection and LS4.C: Adaptation; BOSE, 2012).

The Elevate Learning Progression that lies at the center of the present paper makes use of Wilson's (2009) proposed structure for learning progressions constructed of multiple construct maps. Our multidimensional learning progression consists of Mayr's facts and inferences disaggregated into smaller component dimensions along a horizontal axis (Figure 1) and reflects our analyses of student work collected as part of a pilot study conducted at a high school where the unit natural selection was organized around the facts and inferences (Furtak, 2009). We then developed a construct map for each dimension that illustrates how student understanding can progress from the naïve everyday ideas students may bring to school to the scientifically accurate explanations represented in each dimension of a horizontal axis. Each construct map was then further developed and refined through coding

Dimensions		Construct Maps																					
Fact 1		Fact 2		Fact 3		Inference 1		Fact 4			Fact 5		Inference 2		Inference 3								
Biotic Potential		Population stability		Limited Natural Resources		Struggle for Existence		Transformationist incorrect		Variation		Heritable variation		Differential Survival		Differential Reproduction		Fitness		Speciation		Population changes over long periods of time/deep time	
Construct Maps	Population reproduces but not ideal	Population stability: unclear or vague		Change food source				Environment causes change with genetic basis		Variation: unclear or vague		Heritable: unclear, vague or incorrect		Unclear use of survival		Eugenic reproduction		Survival of the fittest		Speciation: vague or unclear		Population change over time: shorter or no clear duration	
	Biotic potential: unclear or vague	No population stability		Limited natural resources: unclear or vague				Unclear or vague		No variation		No heritability		No survival		Unclear or vague reference to reproduction		Fittest/strongest Survive		No speciation		Species are static	
	No biotic potential			No mention of natural resources				Trait not present		Unclear usage of "adapt to environment"		No heritability		No survival		No reproduction		Unclear use of fitness				No mention of population changing or not changing over time	
																		No fitness					

Figure 1. Multidimensional learning progression for high school students' understanding of natural selection.

TABLE 1
Detailed Information About Random Mutations Construct

Level	Description	Example Student Response
Random mutations	Student describes one or more of the random genetic mechanisms by which new traits arise.	A species changes over time because of random mutations and gene shuffling. Random mutations can cause a change in a species' gene pool. And gene shuffling is the different combinations of genes that come from the parents. If species are separated long enough, the species' gene pool changes.
Environment causes change with genetic basis	Changes occur as a result of genetic mutations in direct response to the environment and/or not random.	Animals mutate to fit in with their natural surroundings. So becoming darker helps to keep them in camouflage.
Unclear or vague	Student refers to mutations or random changes leading to new traits but does not describe a mechanism for how that happens.	If a mutation happens, it can effect the whole species by creating a variety of differences from color change to more or less help against gathering food and protecting against predators.
Trait not present	Description of differences in traits not given at genetic level or denial of change in genes.	I picked my answer because none of the other seemed all the way correct.

students' open-ended responses to assessment items as part of a multiple-year pilot study (see Furtak, Morrison, Iverson, Ross, & Heredia, 2011 for more details on the pilot phase and learning progression development); the final construct maps within each dimension contain between three and five levels. We then developed descriptions of the ideas within each construct map and provided examples of student work.

We focus on several construct maps within our learning progression that collectively represent Mayr's Fact 4, which states that although organisms in a population are on the surface similar, they vary in many ways (Anderson et al., 2002), to illustrate the way that the facts and inferences are broken into smaller component dimensions. On the basis of our analyses of student work and think-aloud interviews, and in line with Darwin's assumption that natural selection acted upon some sort of heritable variation, we disaggregated Fact 4 into three distinct dimensions: (1) Fact 4: Origin of Traits—Genetic Mechanisms/Random Mutations, or the process by which new traits arise as a result of random genetic processes; (2) variation, or the idea that no two individuals within a species or population are exactly alike; and (3) transformationist ideas, or the idea that organisms are not able to transform themselves to adapt to environmental changes. Then, for each of these dimensions, we have articulated the different understandings that students may exhibit before they develop the scientifically accurate understanding represented in the top level, or upper anchor, of each construct map.

We present more detailed information about the Fact 4: Origin of Traits—Genetic Mechanisms/Random Mutations dimension of the learning progression in Table 1. Although the processes by which random genetic mechanisms lead to new traits is not an explicit part of

Mayr's Fact 4, we include it because research has indicated that students frequently have trouble recognizing the role of random molecular processes in science (Ferrari & Chi, 1998; Garvin-Doxas & Klymkowsky, 2008; Odom & Barrow, 1995). The genetic mechanisms dimension (or as teachers came to call it as a shorthand, random mutations) refers to the idea that new traits arise as a result of random genetic processes. These new traits may arise through a variety of processes (e.g., crossing-over, new combinations of genes, mutations), and these new traits may be of many types (helpful, deleterious, or no effect). What is crucial in this category is that students acknowledge that new traits occur entirely spontaneously. The upper anchor involves students stating that random mutations in an organism's DNA or other genetic processes led to the generation of new traits within individuals of a population. The intermediate and lower levels involve students mentioning key words like "random," "mutation," or "DNA" without really explaining those words. Students commonly will fall into this category when they have heard their teacher use these words in describing natural selection, but do not fully integrate these words into a description of random processes leading to new traits, or if they use those words along with anthropomorphic ideas about organisms changing themselves in response to the environment (e.g., the idea that moths can mutate to have the right color to match the bark). The sample student ideas provided in Table 1 are drawn from student responses to open-ended items in the pilot study and are the same as those we provided to the teachers.

The preceding description of the Elevate learning progression for natural selection highlights a number of similarities and differences with other learning progressions, both within and outside the domain of evolution. The term "learning progression" itself has been used to describe a wide variety of representations (Duschl et al., 2011), including those that focus on the development of correct ideas (e.g., those which underlie the NGSS) versus those that build from naïve understandings (Wilson, 2009); those that cover multiple grades (e.g., Gunkel, Covitt, Salinas, & Anderson, 2012) versus smaller units that may be taught at different grade levels (e.g., Alonzo & Steedle, 2009); and those which integrate scientific practices (e.g., Songer & Gotwals, 2012), and those that do not (e.g., Johnson & Tymms, 2011).

Our multidimensional learning progression reflects other learning progressions in this domain that have identified the multiple component concepts that contribute to a full understanding of the mechanism of natural selection. It also integrates a number of common naïve ideas about natural selection, such as the origins of new variations and the role of randomness, that have been identified in previous studies. However, our learning progression also differs from other learning progressions for evolution. First, rather than articulating how student ideas may develop across grades (e.g., Lehrer & Schauble, 2010), it lays out the development of student explanations that could be taught within an instructional unit at one grade level (although the dimensions present could certainly be taught in other units, such as ecology or genetics, or even at different grade levels). In this way, our learning progression addresses a smaller grain size of curriculum as compared to others in this area and therefore it serves the particular purposes of our study. Second, unlike Catley et al. (2005) and Lehrer and Schauble (2012), as well as work in other conceptual domains (e.g., Smith, Carey, & Wiser, 2006), our learning progression focuses on the development of an explanation for natural selection without also taking in practices (to use the language of the NGSS) or habits of mind (as described by Catley et al., 2005).

Our intention in creating a learning progression with this design was multifold. First, we intended that the horizontal sequence of ideas aligned with the facts and inferences could help teachers sequence their existing instructional activities in ways that were faithful to expert explanations of this concept, and that would help them see connections between what Mayr (1982) identified as observable "facts" with the "inferences" that were logical

TABLE 2
Sagebrush High School Teacher Participants

Participant	Bachelor's Degree	Master's Degree	Teacher Credential	Teaching Since
Ann	Philosophy and Biology	ESL and Bilingual Education	Professional, Secondary Science	2004
Carrie	Environmental, Population and Organismic Biology	Curriculum and Design	Professional, secondary science	2006
Nate	Ecology and Evolutionary Biology	Integrated Sciences	Professional, secondary science	2005
Ted	Biology	Curriculum and Instruction	Professional, secondary science	1993

consequences of their preceding facts. We also intended that each individual construct map could act as a resource for teachers in developing formative assessments and other activities, as well as guiding interpretation of student ideas. We hoped that by identifying student ideas at different levels within the construct maps, teachers might be able to provide helpful feedback to move students toward more sophisticated ideas.

Participants

We worked with three partner high schools in the Elevate study from the same district located near a large city in the western United States. At two of the schools, teachers took up the learning progression and DANS without much discussion of these tools, whereas at the third school—we will call it Sagebrush High School—teachers engaged with the learning progression on a much deeper level. Teachers at Sagebrush had a long history of planning assessments collaboratively and had previously participated in workshops around ideal assessment design. Overall the Sagebrush teachers had more frequent and sustained conversations about the learning progression and assessments as compared to the other two schools, and in these conversations they were more transparent in sharing their struggles to understand the representations we were providing. Thus, following case study identification logic (Merriam, 1998; Yin, 2003), we selected Sagebrush for this analysis because because the rich interactions among the teachers illustrated the challenges and benefits of the work in a way that the teachers in the other schools did not. It most clearly illustrated the challenges teachers faced in interpreting the results of assessments linked to learning progressions. The four biology teachers at Sagebrush each taught at least one section of biology during the years of the study, and all were experienced (Table 2; all teachers are identified with pseudonyms).

Pre–Posttest: Daphne Assessment of Natural Selection

We assessed student achievement linked to the learning progression through the DANS, a multiple-choice assessment linked to the Elevate learning progression. Development of the DANS began with administering the Conceptual Inventory of Natural Selection (CINS; Anderson et al., 2002) in its original form to high school biology students as part of a

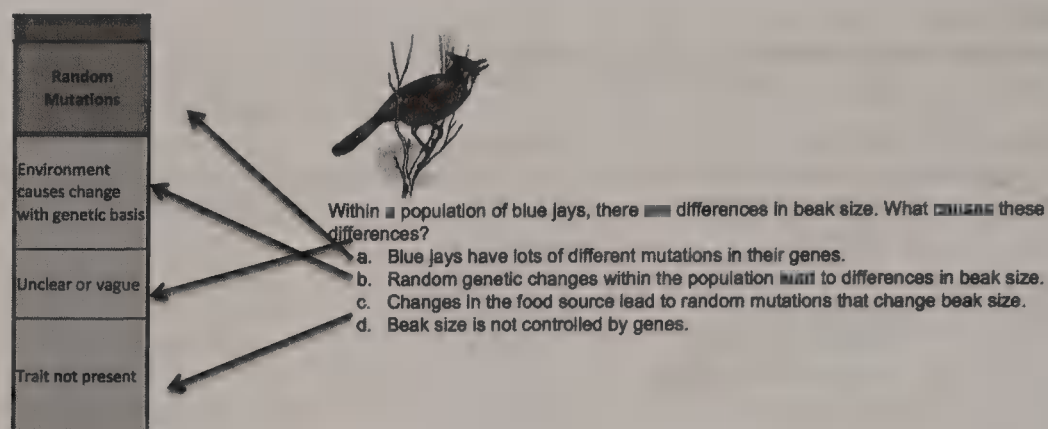


Figure 2. Link between DANS item and Fact 4: Origin of Traits—Genetic Mechanisms/Random Mutations dimension.

pilot study, and over time our experiences in tailoring the CINS items to a high school audience led to the construction of a new assessment in the form of the DANS. Over the course of a 3-year pilot study, we iteratively administered assessments, analyzed student responses, and revised questions to inform and more closely link this assessment to the Elevate learning progression (see Furtak et al., 2011 for more details on the development of the assessment). We performed think-aloud interviews with high school students and matched their transcribed responses with a final set of 22 items to ensure that students were selecting responses based on reasoning processes that we hypothesized. We were restricted to a test with a length that could be administered in one 50-minute class period, and thus we attempted to spread items across the dimensions as much as possible. When multiple items were used to assess a given dimension, we varied the item's context in terms of the organism in which it was framed (Furtak, 2012). Analysis of item performance led us to eliminate items with problematic item-total score correlations, including the items used to assess Fact 2: Stable Population and Inference 2: Differential Survival. This resulted in a reduced set of 17 items common across all testing events. We provide in Figure 2 the link between a sample item and the Fact 4: Origin of Traits—Genetic Mechanisms/Random Mutations dimension to illustrate how, through responses linked to levels on the learning progression, each response option has the potential to provide diagnostic information with respect to student understanding. We further revised the test prior to administration in the third year of the study to improve its measurement properties by adding an additional seven items that addressed Fact 4 and Inference 3 dimensions together, an area of emphasis for teachers in the study, and which boosted internal consistency on the pretest from .31 to .62 (see Alonzo & Steedle, 2009 for a justification of similar ranges of alpha for ordered-multiple choice items with varying contexts).

We administered these items as a pre-post assessment to students enrolled in 10th-grade biology courses at Sagebrush three times: a baseline year (2010–2011) and two subsequent years during which teachers participated in professional development, which is described in the next section of this paper (2011–2012 and 2012–2013). Table 3 lists the number of items by dimension, noting those items that were deleted, retained, and added across the 3 years of the study.

Table 4 provides descriptive information about student performance on the DANS by academic year, and paired samples *t*-tests indicate that students at Sagebrush showed significant pre-post gains each year of the study (2010–2011: $t(305) = -4.827, p < .000$; 2011–2012: $t(301) = -10.066, p < .000$; 2012–2013: $t(203) = -8.399, p < .000$). Table 3

TABLE 3
Number of DANS Items by Dimension

Fact or Inference	Dimension	Baseline and	
		Year 1	Year 2
Fact 1	Biotic Potential	1	1
Fact 2	Stable Population	(1)	0
Fact 3	Limited Resources	1	1
Inference 1	Struggle for Survival ^a		
Fact 4	Origin of Traits: Random Mutations	2	2
	Origin of Traits: Transformationist	3	3
	Variation Within a Population	2	2
Fact 5	Heritability	2	4
Inference 2	Differential Survival	(3)	0
	Differential Reproduction	1 (1)	1
	Fitness	3	3
Fact 4 and Inference 2 ^b		0	7
Inference 3	Speciation	1	1
	Population Changes Over Time	1	1
Total		17 (5)	26

Note: Numbers in parentheses indicate items dropped from the original Year 1–Year 2 DANS.

^aFact 3—Struggle for Survival.

^bSeven items were added in Year 2 that addressed both Fact 4 and Inference 2.

also indicates an increase in pre–posttest effect size between the baseline year (0.48) and Years 1 and 2 of the study (0.79 and 0.67, respectively). We note, however, that conclusions about our intervention and its influence on student learning are not the subject of this paper and will be explored in future analyses.

While we have provided the pre–posttest results here as background, we emphasize that the primary purpose of administering the DANS in the Elevate project has been to focus teachers on interpreting pretest results and using these results to inform areas in which they want to develop new instructional activities and formative assessments; we will describe in the section below the ways in which we supported teachers in engaging in these tasks. The posttest was used primarily for research purposes to track learning gains and to a lesser degree for teachers to identify areas to work on in the future. In this way, our primary intent in administering the DANS was diagnostic and to inform teachers’ instructional practices.

Professional Development

We acknowledge that learning to use a learning progression and interpreting results of an assessment to which it is linked takes time and ongoing support from colleagues, just as with any other ambitious teaching practice. Prior studies have explored how sustained engagement in professional development can help teachers to increasingly attend to student thinking (e.g., Borko, Jacobs, Eiteljorg, & Pittman, 2008; Sherin & Han, 2004). Following a baseline data collection year in 2010–2011, we met monthly on-site with the participating Sagebrush teachers for two academic years (Year 1, 2011–2013; Year 2, 2012–2013) resulting in a total of 21 sessions (10 in 2011–2012, and 11 in 2012–2013).

Our intention was to support teachers in creating formative assessment tools that would elicit a range of student ideas, as well as in learning to listen and attend to a range of student ideas. We also intended to help teachers learn to provide feedback to advance students

TABLE 4
Pre-Posttest Scores for Sagebrush High School

Study Year	Items	N	Pretest				Posttest				Effect Size		
			Minimum	Maximum	Mean	SD	α	Minimum	Maximum	Mean		SD	α
Baseline	17	306	1	14	7.26	2.39	.31	1	17	8.08	2.85	.60	0.34
1	17	302	0	15	6.86	2.38	.40	2	17	8.74	3.13	.65	0.79
2	26	204	3	25	10.63	3.78	.62	1	25	13.17	5.00	.80	0.67

Note: These data include only students who took both the pre- and posttest. The Year 2 N is lower than the previous years due to an overall smaller 10th-grade class.

in their learning. This approach is consistent with the perspective that student knowledge develops over time, and acknowledges the value of connecting to and building upon naïve ideas from students' everyday experiences that do not always map straightforwardly onto the canonical, scientifically accurate descriptions that we ultimately want them to learn. Prior studies have indicated that teachers who hold right-or-wrong visions of student thinking do not enact formative assessment as effectively as those who view student thinking as more of a continuum (Furtak, 2012). Furthermore, the developmental information that formative assessment generates will be most beneficial to teachers if they view student understanding not as binary, "get it or don't" thinking (Otero & Nathan, 2008), but instead as multifaceted, context-dependent, and developing over time.

Our professional development has followed a five-stage, iterative process called the Formative Assessment Development Cycle that we repeated multiple times during the study. The cycle begins by guiding teachers to *explore student ideas* as well their own understandings of natural selection by exploring the learning progression and reports of student performance on the DANS. In the second step, teachers draw on areas they identified from the DANS in which they would like to focus their instruction and *develop tools* such as instructional activities designed to scaffold students' development of a well-articulated explanation for natural selection, as well as formative assessments to elicit more information about these student ideas during instruction. In the third step, teachers *practice using the tools* by rehearsing using the formative assessments together, categorizing samples of student work, and anticipating student responses and the feedback they might provide those students to help them move to more sophisticated levels of understanding. The fourth step has the teachers *enact the tools* during their instructional units and collect student work. Finally, at the end of the school year, teachers *reflect on enactment* by exploring examples of student work, looking for patterns in student responses, and examining student pre–posttest scores relative to the learning progression. The research team guided teachers through each step of the cycle through the course of one school year. Since the baseline year did not involve a professional development component, this paper analyzes two iterations of the cycle. The appendix provides an overview of each of the professional development meetings, their placement in the professional development cycle, and the activities in which teachers engaged in those meetings.

Central to the *explore* and *reflect* steps of the professional development cycle is having teachers analyze student responses to classroom assessments and use the learning progression to interpret those responses and plan instruction accordingly. Reviewing actual test questions with teachers has been established as a worthwhile task for professional development (e.g., Brookhart, 2003); however, due to the fact that the Elevate study is longitudinal for teachers, our funding agency requested we not share the test with our participants to avoid claims that teachers would "teach to the test" if they were familiar with the items it contained. Thus, we needed to develop an approach that would illustrate for teachers how their students had performed on the items without actually showing them the items themselves. As a result of these study constraints, we developed multiple formats over the course of the study to report student results in a way that would support teachers in interpreting student responses relative to the learning progression. We will draw upon the different versions of these assessment reports as artifacts to inform our analysis.

Analytic Approach

The major themes discussed in this paper emerged over a process of several years of iterative enactment of professional development, analysis of pre–posttest results, and reflection upon the process of creating and revising assessment reports. We videotaped

each meeting of the professional development meetings and then created content logs of each meeting in 5-minute intervals. These content logs were then used to create a detailed overview table representing the entire data corpus and which identified the agendas for the meetings, teachers' activities during the meetings, and the representations introduced. Initial analyses of the baseline year data in the spring of 2011 led us to identify issues of across- and with-student variation, a pattern that we tracked in student data through each year of the study, and which informed different versions of the assessment reports described above.

Then, we identified sections of all meetings in which the assessment reports were referenced, whether it was when researchers provided them, teachers were interpreting their results, or using their results to plan their instruction. Each of these segments of the professional development meetings were then transcribed and annotated.

We identified themes from the professional development sessions through both personal reflections on the professional development meetings, both immediately following professional development meetings, as well as retrospectively across the entire study. We then brought evidence from the transcribed professional developments to bear on the interrogated themes we had identified. We then iteratively discussed these themes, confirming and disconfirming evidence for these themes, created research memos, and revised the themes accordingly.

We established validity and reliability in this approach by making explicit our biases and assumptions with each other, checking our claims with each other, and triangulating our claims from multiple professional development meetings collected over the 2 years of professional development enactment (Merriam, 1998). Validity was further reinforced by the fact that the third author did not participate in the planning or enactment of the professional development meetings and did not know the teachers. Ultimately, we agreed upon a set of themes supported by evidence from the pre-posttests and professional development videos. We created the following case report arranged around those themes and supported with evidence across the 2 years of the study.

POTENTIAL AND CHALLENGES FOR TEACHER USE OF LEARNING PROGRESSIONS AND CLASSROOM ASSESSMENTS

In this section, we draw upon Sagebrush students' responses to the DANS as well as videotapes of professional development meetings with the Sagebrush teachers to illustrate the challenges we have encountered in our efforts to support teachers in interpreting the DANS relative to the Elevate learning progression. In the sections below, we divide these challenges into three sections: helping teachers understand the learning progression, using the learning progression to interpret assessment results, and using the information these tools contain to plan next steps for instruction. We connect each area we identified with previous research and illustrate each with examples from our study. By drawing upon data from multiple years of the study, we provide a historical perspective in terms of teachers' reactions and responses to the various tools we provided to support them during the study.

Understanding the Learning Progression

The first step in being able to match results of classroom assessments to the learning progression is for teachers to be able to read the learning progression and understand the information it contains. Student growth along the levels in a learning progression may follow a number of different patterns (e.g., Corcoran et al. 2009; NRC, 2007) and is not developmentally inevitable, but occurs in the presence of a particular set of learning

experiences (Stevens, Delgado, & Krajcik, 2010); nevertheless, the most common format for a learning progression is linear and hierarchical, as is our own learning progression for natural selection. Representations such as these might reinforce teachers' impressions that learning proceeds in only one path, or may furthermore reinforce traditional "right-wrong" ideas about student learning (Furtak, 2012).

Teachers at Sagebrush often asked questions about what the different levels of the learning progression meant and tried to map the levels onto their understanding of what the "correct" answers were. For example, the first time teachers were introduced to the learning progression at the October 2011 meeting, Nate asked whether "the highlighted one is the correct one?" He later rephrased the researchers' descriptions of the different levels by summarizing, "So was it like correct (pointing to top level), could be correct (middle level), and then the unclear vague is just incorrect (lower level)?" In these instances, Nate seemed to be trying to align the levels on the learning progression with his perception of what was correct and what was wrong. This view of student ideas persisted through the first year of the professional development, not only when teachers were looking at the learning progression but also when they were planning formative assessments and other instructional activities. For example, in the April 2012 meeting, teachers used words and phrases like "emphasize," "break down that misconception," and "make sure they understand" when discussing how they intended to enact the activity. These words and phrases suggested teachers were viewing student ideas as needing to be broken down and that the right answer needed to be stressed, indicating that they were still focused on correct and incorrect interpretations of student ideas.

In a way, this meant that teachers and researchers were often speaking on parallel planes about the information the learning progression contained. Teachers focused on whether particular ideas were correct or incorrect, and researchers tried to help the teachers reframe their interpretations to acknowledge the multiple aspects of student thinking it represented. The following excerpt from December 2012 illustrates that teachers continued to interpret the student ideas represented in the learning progression during the second year of their participation in the professional development. In this interaction, Ted talks with Deb and Erin, the researchers and facilitators of the professional development, about the number of students that scored at the lower levels of the learning progression:

- Ted: ... I would say that my students generally speaking did not understand the concept.
- Deb: It's not necessarily that they didn't understand that concept, it's that they had a particular idea they brought in.
- Ted: Right, they didn't have the correct understanding.
- Erin: Their everyday experience was pointing to this instead of the scientifically accurate answer.

This excerpt illustrates the contrast between how Ted discussed student ideas as being correct or incorrect and how the researchers framed the ideas in the lower levels as the everyday ideas students had prior to instruction.

The researchers struggled to push beyond teachers' correct/incorrect interpretations and to help them identify a range of more and less sophisticated student ideas as represented in the learning progression. During the March 2013 meeting, Nate tried to describe his impression of the levels of the learning progression when he said, "Is each level going down, is that a different part of that fact or is that like a dumber and dumber answer?" In this instance, Nate seemed to better understand by phrasing the levels as increasingly "dumber" as they went down the chart. Following this exchange, Ted also stated that, if the levels really indicated different types of understanding, it would be helpful to include a

blank or “nothing” level on the learning progression at the bottom, as that was “what kids are thinking when they walk into the class.”

In the excerpts above, teachers were interested in understanding where the correct answers were represented, and then used their own language to describe the lower levels on the progression, such as “incorrect” or “dumber.” Furthermore, Ted also indicated that the existence of a blank level on the learning progression might better represent the way his students were coming to class—in effect, a *tabula rasa* view of student learning. These instances suggest that while Nate was beginning to note different levels of understanding within the categories, Ted still appeared to hold a dichotomous, right/wrong perception of student ideas.

During the last meetings of the year in May of 2013, however, all four of the teachers began to show more attention to the developmental information in the learning progression. During the first meeting that month, teachers brought samples of student responses to the formative assessments and read through them, using the learning progression to identify different ideas in student responses. Teachers talked about student responses indicating that certain beak types that had gone extinct in their model could come back if the conditions were right, reflecting the idea from the learning progression that the environment could compel particular types of beaks to arise. In the excerpt below, Carrie reads through a student response and interprets it:

- Carrie: This is a kid who actually tries very hard [reading from student work]. “It could come back because the birds need to change in order to survive in the environment. Brand new birds will be made.” So it sounds [as if the beak coming back could be] possible. Completely-
- Deb: – So he’s conflating some kind of reproduction with like the idea that new birds, as in like changes the traits, or -
- Carrie: - new birds need to be made, so of course they’re going to be made to fit that environment . . . so that’s where that weird flip flop happens where they no longer equate it to randomness, they go, well if that environment’s there, then a bird will be made to fit that environment, even though the environment selects for it. That’s like where their logic actually, so like this [holds up student work] is exactly it, and so I have a few more just like that, but there are only three out of this pile that talks about it. Last year it was like the whole pile. So that’s really good.

In this interpretation of the student response, Carrie takes a sample student response apart and identifies a variety of ideas it contains in terms of the role of randomness, fitness, and selection. She does not call the response “right” or “wrong,” but rather focuses on nuances in the student thinking. Such discussions did not focus as much as previous meetings on student misconceptions, but focused more on the difference between student ideas that frame change on the level of populations versus individuals, and different possible sources for less sophisticated student ideas.

Interpreting Assessment Results Relative to the Learning Progression

Our main objective in engaging teachers with the learning progression was for them to use it as an interpretive framework for results of classroom assessments. That is, we wanted teachers to be able to make valid inferences about what students know based upon the information contained in the learning progression and accompanying reports of student responses to the DANS. Yet from the start—in fact, due to the very nature of assessment of

	Single Dimension	Whole Learning Progression
Across students	All students' responses to the different items within the same dimension	All students' responses to all items, for all dimensions within the learning progression
Within students	One student's response to different items within the same dimension	One student's response to all items, for all dimensions in the learning progression

Figure 3. Four levels of complexity in interpreting diagnostic information from a multidimensional learning progression.

student thinking—the information the test provided was complex and led to difficulties in helping teachers to interpret that information relative to the learning progression.

We identify four different types of complexity in the results of student responses to the DANS, illustrated in Figure 3. For example, when looking across students, we would expect that students within a class would show different levels of performance on a dimension (across-student dimension complexity, as would be represented as cross sections within each dimension for all students), as well as different patterns of performance on an entire construct from one student to the next (across-student construct complexity). We will present our results using the within student and across student lens and within each we will address the single dimension and whole learning progression factors that arose in the course of our work. In addition, a student may not respond in the same way to all assessment items linked to a particular dimension (within-student dimension complexity, as illustrated by looking at a single student's response between items in a single dimension). This fact has been acknowledged by assessment developers in the biological sciences, who have found that the organism or context of items may influence the way students respond to it (Duncan & Hmelo-Silver, 2009; Heredia, Furtak, & Morrison, 2012; Nehm & Ha, 2011). Similarly, students may respond at different levels of sophistication on the different dimensions of a complex learning progression (within-student construct complexity, as illustrated by looking at a single student's response pattern across all items and all dimensions), as found by Steedle and Shavelson (2009).

Representing Across-Student Complexity. We intended the assessment reports to represent not only variation in the ways students responded to items for a single dimension or across the whole learning progression but also variation across the whole class. In the version of the assessment report format we created in the first year of the study, we aggregated student responses to distractors at different levels and presented the data as a set of tables for teachers. Figure 4 shows a sample table with student responses to one item about Fact 4: Origin of Traits—Genetic Mechanisms/Random Mutations leading to new traits in animals (student responses to random mutations items were disaggregated according to items framed as plants and animals, whereas other tables presented means of student responses within dimensions). This form of the assessment report showed the scientifically accurate level of each dimension of the learning progression (in the darkest shaded row of the table) and the lower levels of the learning progression in the left column. We then included the percentage of students selecting response options corresponding with each level and disaggregated results for teachers by class type. In this way, we intended to show teachers the percentage of students with different ideas prior to

			Class Type		Total
			General Biology	Honors Biology	
Fact 4 Origin of Traits <i>Animal</i>	Random Mutations	Count	43	14	57
		Percentage within Class Type	41.3	60.9	44.9
	Environment causes change with genetic basis	Count	28	5	33
		Percentage within class type	26.9	21.7	26.0
	Unclear or vague	Count	22	4	26
		Percentage within class type	21.2	17.4	20.5
	No genetic basis	Count	11	0	11
		Percentage within class type	10.6	.0	8.7
Total	Count	104	23	127	
	Percentage within class type	100.0	100.0	100.0	

Figure 4. Sample pretest assessment report excerpt from fall 2011.

instruction, as well as to help them track student progress following instruction. This approach was intended to help teachers glean more information about student understanding than they would get through a simple percentage of correct items aggregated across multiple dimensions.

In fall 2011, the second year of the study, we used this assessment report format to present pre–post data from the baseline year to teachers, along with pretest data from the current year. Given the number of tables—at least one for every dimension of the learning progression—each teacher in the study received a multiple-page packet.

In the professional development meetings at Sagebrush in the fall of 2011, when we were using this initial form of the assessment report, teachers used the data to identify whole-class trends across different dimensions of the learning progression. One such moment is illustrated in the following excerpt, taken from the October 2011 meeting, in which teachers explored the pre–posttest data from the baseline study year (2010–2011), as well as the pretest data from 2011:

Nate: [looking at his assessment report] It’s like they did the best on the variation of animals and then, then limited resources in my classes

Erin: They did the best like highest percentages of right answers?

Nate: Right.

Erin: Okay.

Ian (research assistant): So that’s, that’s where we want to go with this eventually is to have folks kind of identifying ones where they were strong also ones where then you see . . .

Erin: Yeah, so what else . . . Ted you were going to say something?

Ted: Oh looking at this, the concept that I think that I noticed that my students struggled with the most, I think, was mostly in Fact 4, random mutation or not random mutation. That was a little low I thought, lower than I perhaps should like because it’s such a general not general but a fundamental concept . . .

This excerpt illustrates the teachers using the assessment reports to note what students understood, what they struggled with, and how this compared to their expectations. The fact that teachers were able in this meeting and in other instances to observe trends across the class and match them with information in the learning progression reflects utility in the assessment–learning progression link. However, we also noted that while our initial version of the assessment report format did identify whole-class data, it did not represent

within-student variation, and teachers did not make student-level observations on the basis of these reports.

Representing Within-Student Complexity. One challenge in helping teachers interpret the assessment results was that students did not always respond similarly to items within the same dimension, and likewise did not respond at the same level across the different dimensions of the test. While part of this issue is associated with measurement error, we also have found that it can be attributed in part to the context of the items, such as whether the item referenced an animal or a plant (see Furtak, 2012, for a more detailed explanation of this result in and of itself, and Alonzo & Steedle, 2009, for another example of variation in student responses according to item context). Indeed, as we noted above, we presented student responses to items about plants versus animals linked to the Fact 4: Origin of Traits—Genetic Mechanisms/Random Mutations dimension separately on the assessment reports to highlight this variability.

The fact that our learning progression is multidimensional generated additional levels of complexity for teachers in looking at student responses for the whole class. Prior studies have indicated that it is not always possible to locate students at a single level on a unidimensional learning progression (Steedle & Shavelson, 2009). Similarly, Corcoran et al. (2009) acknowledged that “There may be multiple possible paths and progress is not necessarily linear.” (p. 38). When looking at a learning progression that involves multiple dimensions, then, we would anticipate that students might not respond similarly to items linked to different dimensions, but would rather illustrate different levels of proficiency on different dimensions of the learning progression.

To illustrate the multiple profiles of proficiency that a multidimensional learning progression might capture, we show in Figure 5 a representation of two Sagebrush students’ responses to the 17 retained DANS items in the fall 2010 pretest. The figure shows student responses in terms of their location on the construct ranging from most to least sophisticated, from top to bottom. The items within each dimension are shown across the top.

Both students had the same score of 7 on the 17 items 2011–2012 pretest. However, as the figure illustrates, their responses to the items varied across the subset of the dimensions of the learning progression tested on this version of the DANS. This figure illustrates that students do not score consistently across the different subdimensions of the learning progression. In other words, it is clear that a student can appear to be very sophisticated in some dimensions and very naive in other dimensions. In addition, this phenomenon appears to hold true for students with high overall scores as well as low overall scores. Both students appear to have a high degree of variability in their scores on individual items. It is also interesting to look at the student responses in Figure 5 and attempt to reconcile that both students would have received the same score on this assessment given that their responses are distributed so differently across the dimensions on the learning progression.

On the basis of our experience that teachers made only class-level inferences from our initial version of the assessment reports, as well as our analyses of assessment data as shown in Figure 5 (which we did not use with the teachers, but created as part of our own analytic process), we searched for a format that would better support teachers in noting that students did not always respond the same way to items linked to the same dimension of the learning progression. We drew upon Minstrell et al.’s (2008) Teacher Report that breaks down student responses to questions from the DIAGNOSER software program broken down by facet level, similar to the levels on a learning progression. Minstrell et al.’s Teacher Report also shows the percentage of students within a class that responded at least once at a particular level, or at least twice—thereby indicating, to some extent, the consistency

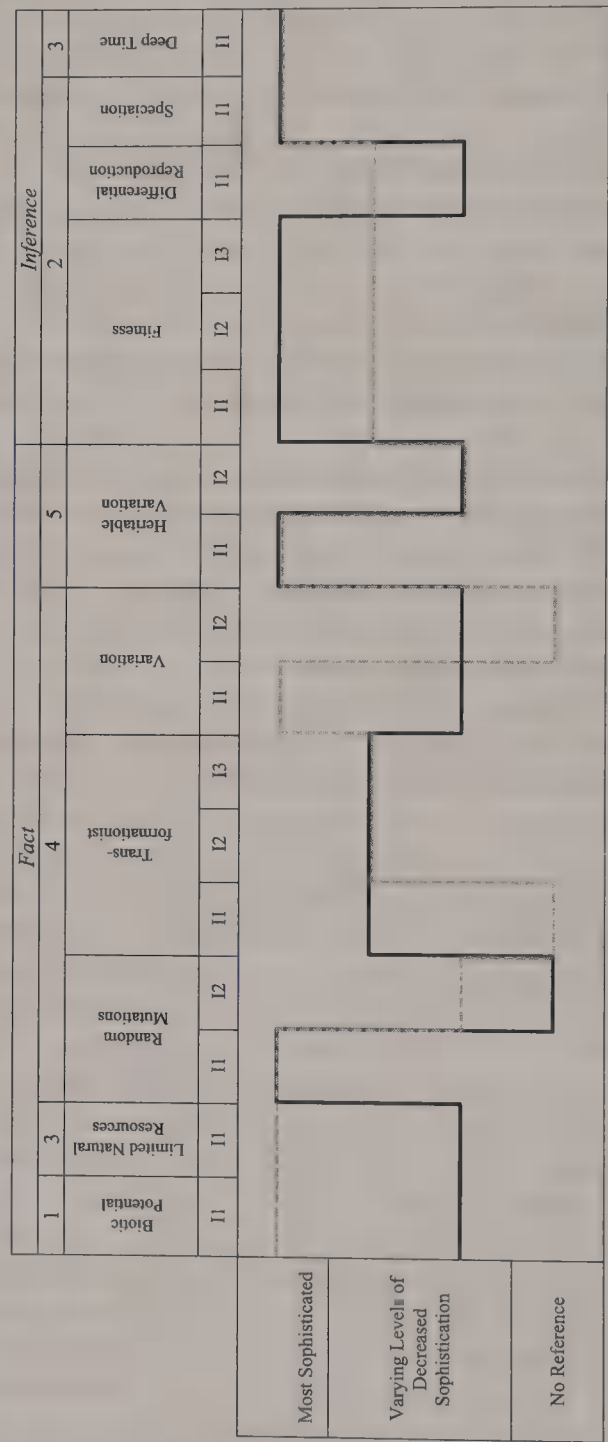


Figure 5. Performance of two students on each item of the DANS, by dimension.

Fact 4: Origin of Traits (two items)

Level	At Least Once	At Least Twice	What Students Need to Work on	Next Steps
<i>Genetic changes happen through random genetic processes</i>	58	17		
Genetic change without mechanism	62	20	Students acknowledge that there is a genetic basis to change, but do not specify what that is	Have students explore mechanisms for genetic change (e.g., crossing over, recombination of genes, mutations, etc.) and related those changes in genotype to changes in phenotype
Environment causes genetic changes	39	7	Students understand that changes in traits has a genetic basis, but they think that change is caused by the environment	Students need opportunities to explore how the environment cannot generate changes in DNA
Trait change has no genetic basis	17	1	Students do not relate changes in traits to genes	Review DNA-RNA-protein-phenotype link for students, help them to see that phenotype is controlled by genes

Figure 6. Sample pretest assessment report excerpt from fall 2012.

within student's responses. Minstrell has also explored adding additional information about what students need to work on and what next steps teachers might take in their instruction to the Teacher Report (Minstrell, personal communication, 2012), a step we also adopted in revising assessment reports for the 2012–2013 school year. Our revised assessment reports used in the 2012–2013 school year drew upon Minstrell et al.'s (2008) format, with the important difference that we reported results relative to the ordered levels of the learning progression, rather than goal and problematic facets. To simplify the amount of information provided to teachers, we also aggregated the categories of the learning progression into fewer sections to focus on the parts of the learning progression that teachers had taken up in the study and showed the number of students who had responded at each level one or two times (Figure 6).

When we shared this format of the assessment report with teachers at Sagebrush for the first time in December of 2012, the discussion immediately went to interpreting what the numbers in the columns meant. If they were not percentages, did that mean that students could be double-counted? What if there were three versus four items? The research team and teachers went through how the numbers had been calculated to the extent that the "what students need to work on" and "next steps" columns were not discussed as deeply. The following excerpt from this meeting illustrates Nate's struggle to understand the numbers in the columns of the assessment report:

Nate: So then more, approximately half of mine answered the question correctly at least once, but that small percentage answered it correctly twice.

Erin: Right, so it's like harder to get, the students who understand more.

Nate: I get that but what I don't get is how do the rest of the numbers add up.

Erin: Well it will be a similar pattern.

Nate: 'cause if they don't answer that question correctly then they fall in the incorrect category right?

Erin: Yes, and there is no incorrect.

Nate: Aren't these the incorrect categories?

In the preceding excerpt, Nate not only was unsure of what the numbers meant on the report, but was also trying to map the information contained in the learning progression onto his traditional right/wrong framing of student thinking. This conversation occurred during the same time in which teachers were still grappling with their dichotomous view of student thinking.

However, the revised assessment report did help teachers focus on the fact that students do not reliably respond to items within a particular dimension of the learning progression in similar ways, as discussed in the section above. This is evidenced by Nate's response to the report, when he stated, "I don't understand how they could think it was right, I mean have the right concept once and then so often screw it up." This comment prompted Ted to refer to his own understanding of how trustworthy the items were, although he did not describe it as such:

Here's the thing though, if you think about it, you know, remember when we first started writing tests and talked about having, you know, a standard has to have at least five questions. It's kind of the same idea you've got to have as much—so in reality I would say that these questions here that only have two items really you can't get a—aren't really valid in terms of trying to find data to accurately analyze. Where when I look at this one that has six items, I'm like, okay wait a second now, we have, we can get—as we're slowly working through how to read this.

Here Ted refers to his own understanding of how many items are needed to make reliable judgments about a single concept, which he had learned previously through other professional developments at his school. This understanding of reliability in a classroom context is extremely relevant to the assessment, as only a few items were used for some dimensions. Nate responded again, though, returning to his frustration that multiple students got some items right and others wrong within that same dimension:

But what does that even tell you? I mean what does that even mean to you though? So of those six items almost all of my students got at least one of them right but then almost all of them got that first one wrong so that was the one that they struggled with. Is that where we're going with this? Like almost all of them got one option, one selection right, but then right below that almost all of them chose that wrong option at least one time so that's where the wrong choices are going? Right? That's where their eyes are going or that's where their minds are going, towards that wrong selection.

Here Nate again struggles to make sense of the fact that although most of his students got one question right on the dimension, most of his students also picked alternative—not scientifically accurate—responses at least once.

In the spring of 2013, when we provided teachers with students' posttest results in the same assessment report format, teachers spent less time interpreting the report format and more time focusing upon how students had performed between the pre- and posttests that academic year. Teachers noted how students had moved between levels from pre to post, noting when students had moved from middle levels of the learning progression to higher levels, or—in some cases—to lower levels. This version of the assessment report also facilitated teachers in looking at patterns across the dimensions of the learning progression. When Ted was looking at the transformationist dimension, he took a moment to reflect on his students' responses:

I can already be able to translate this [assessment report]. Some of the kids . . . so they looked at that answer and said, "No people don't change to fit into the environment," all

right, but when you said that traits change due to a random genetic process they thought that was true. But then the environment causes genetic changes, all right, so in their head are they thinking ahead saying well technically the environment does cause a genetic change because the population dies out and there's a shift . . .

In this interpretation, Ted identified how his students responded at the top level on the transformationist dimension of the learning progression, but at lower levels on the Fact 4: Origin of Traits—Genetic Mechanisms/Random Mutations dimension. His response illustrates how, toward the end of the 2013 school year, Ted was able to draw upon the revised version of the assessment report to note subtle differences in student response patterns within the different dimensions of the learning progression.

Using the Learning Progression and Assessment Results to Inform Instruction

We acknowledge that the amount of information the assessment reports provided teachers might have been paralyzing in the absence of professional development; nevertheless, we were determined to create assessment reports that would support teachers in identifying next instructional steps based upon their interpretations of data about what their students knew. We thus structured our meetings to support teachers in drawing upon information from the reports and learning progression to plan formative assessments and other instructional activities based on those results.

During the first year of professional development meetings, we asked teachers to carefully analyze assessment reports and determine pieces of the learning progression on which they wanted to focus. Our intent was for them to then spend the academic year planning formative assessments and instructional experiences to support students in learning the concepts represented in those pieces of the learning progression. While a large part of the conversation in the October 2012 meeting consisted of looking at trends in the assessment reports and selecting pieces to focus upon, teachers also engaged in a conversation about the source of ideas that were revealed in the assessment reports. In the following exchange, Carrie and Ann noted that although their students seemed to understand that it takes a long time for new species to form, at the same time the students appeared to be confused about whether organisms are able to change themselves:

Carrie: . . . They know it takes a long time, it is a hard concept to understand, and why, if they can know it takes a long time and get it right on the test, it doesn't mean that they understand all the principles behind why, so it's like there's all these things that don't line up.

Ann: I think some confusion comes too because the kids learn about all these cool animals that can change, you know, like camouflage or whatever and they can't be seen, and I think they take that as that they're then changing to adapt to their environment, right? So it's a lot of things for kids to sort out, that that's actually just a genetic phenomenon for that particular species, you know? That they still have a preset genetic makeup, but the fact that they can change color or blend in, or you know make that adjustment like the lizard family people . . . it's hard for kids, because they know that knowledge from watching things . . .

This exchange illustrates the potential of the assessment reports when compared with the learning progression, in that they not only helped teachers identify patterns in what students understood and what challenged them but also provided opportunities for teachers to have conversations about students' prior experiences that might have led them to those ideas.

During the fall and spring of the 2011–2012 school year, teachers read closely through their assessment reports to identify dimensions in the learning progression with which their students had struggled, not only in the pre–posttests from the 2010–2011 baseline year, but also by looking at the 2011 pretest reports:

- Nate: My two lowest were both in fact 4 the origin of traits plants and the transformationist misconceptions.
- Carrie: I had the same except for one is speciation was really low as well. But yeah, those are the lowest two.
- Ted: Well, we should focus on speciation.
- Nate: We're not looking at -
- Carrie: - based on . . . the random mutation . . . is the thing that they struggle with, and the transformationist.
- Nate: Oh I didn't see the back, speciation is my lowest.

Ultimately, the group decided to focus on the Fact 4 dimensions of speciation, origin of traits and transformationist misconceptions, and changes over time. Then, at subsequent meetings, the teachers brought their previous year's lesson plans and activities to locate the areas in their unit in which these concepts were taught. They identified three activities—a model of variation and selection in which students engage in a limbo activity, a simulation of Peter and Rosemary Grant's study of small ground finches in the Galapagos Islands (Grant, 1986), and a video about cichlids living in Lake Tanganyika in the African Rift Valley—that they wanted to revise and use to integrate formative assessments.

As the unit approached in February, March, and April 2012, teachers revisited each of these activities to rework them based upon what they had learned in the assessment reports. In the March 2012 meeting, the group discussed how an activity in which students pick up beans and seeds using utensils might actually have led to their struggles in understanding variation and speciation, as follows:

- Nate: . . . the four different beak types you know the four: the spoon, fork, the knife beak and test tube beak—I can see them in their minds thinking these are different types of birds-
- Ted: - these are different species -
- Nate: - not that this is one species of finch with different, with variation in beak type.
- Ian: Well and I can see that . . . what led to that I think is too is the fact that we are talking about four different, very different utensils.
- Nate: Yeah.
- Ted: And in the actual study we are talking about just millimeters of difference between beak types you know
- Erin: Yeah.
- Ted: Yeah, . . . to do something like that you know you could use a big spoon, small spoon, and a you know like those little spoons, sample spoons . . .

This excerpt was drawn from an extended conversation in which all four of the teachers participated in problematizing the way beak size was represented in a model that the department had been using with students for more than 10 years. The subsequent meeting in April 2012, immediately prior to teaching the unit, involved the teachers rehearsing with each other how they intended to introduce ideas of variation and the model of using spoons of different sizes to represent small variations in beak size within the population of medium ground finches on the Galapagos.

The next year of the study, the 2012–13 school year, involved teachers noting that they thought this change had positively impacted student responses on classroom activities, and

they turned their attention to revising the limbo activity to focus more on representing and discussing variations in height among members of their classes by creating a whole-class histogram, making predictions about success at limbo, and then speaking about possible consequences of these variations on survival given a particular scenario in which food was kept in a cave with a low entrance. The preceding examples illustrate the ways in which teachers drew upon results from assessment reports to inform changes in their instruction.

DISCUSSION

In this paper, we have reported on a multiple-year study in which we have engaged teachers in ongoing professional development to help us understand the ways in which teachers' understanding of learning progressions develops and unfolds over time. We have explored the potential utility as well as challenges we have encountered in using learning progressions and classroom assessments with practicing teachers. In this section, we briefly summarize each category of the findings of our study and discuss their implications for research in this area. We then frame the implications of our findings for future research in the area of linking learning progressions and classroom assessments.

Understanding the Learning Progression

Analysis of our professional development meetings revealed a number of instances early in the study in which teachers struggled to make sense of the learning progression relative to their own, more dichotomous views of student thinking. Over time, however, we found that teachers were increasingly able to disaggregate student responses into different categories that mapped onto the learning progression. These results indicate the importance of supporting teachers in professional development settings in interpreting the learning progression, in eliciting their often traditional views of student thinking, and supporting them in identifying the developmental affordances of the way student ideas are represented in the learning progression.

We acknowledge that the way student ideas were represented in our learning progression may have unintentionally reinforced teachers' dichotomous views of student ideas. For example, many of our dimensions had as their lowest anchor a level that indicated a particular student idea was "not present," which may have reinforced teachers in thinking that students came to class with no prior ideas about these dimensions. Furthermore, while Mayr's (1982) facts and inferences helped to disaggregate some of the complexity and structure of the expert view of natural selection and identify the many elements that need to be woven together to construct a well-articulated explanation, they did not allow us to portray the complexity of the ideas students bring to class. The learning progressions constructed by Catley et al. (2005), Lehrer and Schauble (2012), and Metz (2010) are more successful in representing the "seeds of evolutionary thinking" and how these ideas merge and change as students progress through school. In addition, while our decision to disaggregate particular student misconceptions into a "transformationist" construct may have helped teachers to focus exclusively on this idea, it failed to capture how these ideas manifest themselves across a number of dimensions of natural selection, as illustrated by Shtulman (2006).

In retrospect, we realize that there are also numerous ways in which we might have better supported teachers in using the learning progression to interpret student ideas in concert with a learning progression that better captured—in a positive sense—the way that student naïve ideas build into scientifically accurate ideas. While we always attempted to frame particular student ideas, such as the ideas that individual organisms change in response

to the environment, as productive for learning, we acknowledge that this productivity was not reflected in the design of our learning progression. A different representation might have helped us reach this productive perspective with teachers earlier in the study, rather than struggling for at least a year to overcome teachers' tendency to interpret the learning progression as suggesting they be used to "find and replace" students' naïve ideas with scientifically accurate conceptions. This was certainly not the intent of the Elevate project, and yet it may be an unintended outcome of particular learning progression designs. In addition, alternate approaches to representing student ideas, such as Minstrell et al.'s (2008) approach of organizing ideas into goal and problematic facet statements might help teachers categorize and identify student ideas without implying particular developmental sequences.

Future research may explore how different types of representations of student ideas in learning progressions might support more—or less—reform-oriented thinking about student ideas on the part of teachers. If we intend to realize the potential of learning progressions as tools to support instruction, it will be essential to study the constraints and affordances of different types of learning progressions in supporting teachers' developmental thinking about student ideas. Furthermore, explicitly embedding professional development about learning progressions in theoretical perspectives about student learning may further support teachers' development. Gunckel (2013) has performed early work in this area by identifying the pedagogical content knowledge necessary for teachers to use learning progressions, and further work in this area can help the field to better understand the most effective supports for teachers.

A revised version of our learning progression might move away from a Mayr's (1982) decomposition of an expert view of natural selection and instead focus on the way ideas about natural selection develop across grade level. For example, ideas about variation might be introduced to students first, then looking at similarities among related organisms, and then ideas about heredity. In this way, a different form of learning progression—similar to those of Lehrer and Schauble (2012) and Metz et al. (2012)—would have as its overarching structure a progression of how ideas develop over time and would better represent the productivity of particular student ideas. This approach would draw upon a principle articulated by Wiser, Smith, Doubler, and Absell-Clarke (2012), who intended their learning progression to establish "learning goals in terms of anchors and stepping stones rather than in terms of pieces of expert understanding" (p. 3). In this way, an improved learning progression would more explicitly integrate useful intermediate models that, although not yet expert, may have some coherence and help in moving student thinking forward.

Interpreting Assessment Results Relative to the Learning Progression

We have found that different forms of reporting formats for assessment results relative to learning progressions focus teachers on different levels of complexity, either looking across entire classes or focusing on within-student response patterns. We found that the assessment report format we used during 2011–2012 did help teachers to focus on whole-class trends and to identify areas of the learning progression to focus upon, but at the same time it did not reveal how individual students were responding within dimensions of the learning progression. The version of the learning progression from 2012–2013 better represented consistencies and inconsistencies in student response patterns within dimensions and helped teachers to see that just because students responded to a question at a particular level once, they might not do so multiple times.

We also note that there may be utility in breaking multidimensional learning progressions into smaller component pieces and smaller accompanying assessments. In this case, teachers might work with shorter tests closer to the times in which they might be using the results of those tests; in a sense, cascading sets of diagnostic assessments that would be used on a regular basis to inform instruction. According to Briggs, Alonzo, Schwab, and Wilson (2006), such assessments could contain as few as eight ordered multiple-choice items linked to dimensions of the learning progression and still have reliable results reflective of student thinking.

Using the Learning Progression and Assessment Results to Inform Instruction

Our results indicate that teachers were able to draw upon information from the learning progression and assessment reports to identify areas of their instruction for improvement. During each year of the study, the teachers participated in the ongoing process of investigating the results of the pretest and then slowly, over a period of multiple meetings, reflecting upon their current instruction and making targeted changes intended to improve student learning along certain dimensions of the learning progression. In this study, we viewed teachers as authors of instructional materials and not merely implementers, honoring their rich previous experiences, knowledge base, and what they may bring to instruction. If the potential utility of assessments linked to learning progressions is to be fully realized, teachers need to be active participants in taking action upon the information that these representations contain.

We note that the teachers at Sagebrush High were fortunate to have the flexibility to reorganize the curriculum materials they used in ways that they found to be more consistent with the learning progression. It is possible that other schools with less curricular freedom might be more constrained in their efforts to improve instruction relative to the learning progression. Furthermore, we recognize a trade-off in terms of the timing of administering preassessments and unit redesign. Clearly, if teachers receive the results of preassessments too close to the start of an instructional unit, they will not have time to reflect and plan for changes in response. However, the more distal the administration of a preassessment, the less urgent the process of design may appear to teachers. Some type of support is necessary to allow teachers the time and opportunity to adequately draw upon the results of assessments linked to learning progressions to inform their instruction.

CONCLUSION

We believe that our study is representative of the struggles researchers in this area have been and will be meeting in the course of their work, and advise those working in the area of learning progressions and classroom assessment to be aware of the challenging nature of this work. Learning progressions themselves are complex tools, and researchers must address the ways in which teachers employ information about how students learn, and how assessment should be used for classroom or formative purposes. As our results illustrate, teachers' understandings of how students learn are at play in these instances in which they encounter the learning progression and classroom assessment. As assessments linked to learning progressions are rolled out in more classrooms, the assessment designers will need to be aware of the realities of the schools in which they are working and may be forced to make compromises in the length or depth of the assessments they are using to make them short enough to not be too demanding of instructional time.

Our results also clearly indicate that tools alone will not help teachers realize shifts in practice. Sustained professional development that carefully attends to teachers’ understandings of student thinking and how students learn will be necessarily to realize the potential that these tools offer. Researchers engaged in developing learning progressions and assessments linked to them should carefully consider the types of supports they are providing teachers to interpret these results. The design of professional development to better support teachers will be crucial in realizing the potential that learning progressions and classroom assessments have to inform instructional practice.

APPENDIX: SUMMARY OF PROFESSIONAL DEVELOPMENT MEETINGS

Session	Date	Phase	Activities
1	8/18/11		<ul style="list-style-type: none">▪ Discuss Elevate project with researchers and become acquainted with project goals• Set meeting and organizational norms
2	9/28/11	Explore student ideas	<ul style="list-style-type: none">▪ Discuss purposes of professional/teacher learning communities▪ Introduce teachers to Elevate learning progression▪ Problematize traditional approaches to assessment and contrast with design of Elevate• Distribute and begin discussion of 2010–2011 assessment reports
3	10/26/11	Explore student ideas	<ul style="list-style-type: none">▪ Discuss conceptions of/approaches for formative assessment▪ Teachers deepen their understanding of the content in the learning progression▪ Teachers share interesting trends/areas of concern from 2010–2011 assessment reports▪ Teachers explore 2011 preassessment reports and compare them with those from 2010–2011▪ As a group, identify two to three areas of learning progression as areas of focus for formative assessment this year
4	12/7/11	Develop tools	<ul style="list-style-type: none">• Discuss general approaches to formative assessment• Come to a better understanding of the ideas represented in the areas of focus on learning progression• Begin brainstorming ideas for common formative assessment

(Continued)

APPENDIX: CONTINUED

Session	Date	Phase	Activities
5	1/24/12	Develop tools	<ul style="list-style-type: none">• Discuss teachers' views of how students learn• Invite teachers to sequence the dimensions of the learning progression for themselves, and discuss their rationales for so doing• Discuss learning progressions and their affordances and limitations of them in supporting instruction• Develop a first-draft formative assessment linked to the Fact 4 dimension
6	2/29/12	Develop tools, practice using tools	<ul style="list-style-type: none">• Draw on materials, assessment results, the learning progression, and teachers' own knowledge of natural selection to develop formative assessments for use in the area of natural selection they identified from the assessment reports• Discuss different teaching approaches/strategies for enacting the formative assessments• Anticipate likely student responses to the formative assessments
7	3/21/12	Develop tools, practice using tools	<ul style="list-style-type: none">• Revise draft formative assessments• Anticipate student responses and teaching moves
8	4/18/12	Develop tools, practice using tools	<ul style="list-style-type: none">• Finish designing formative assessments, anticipate student thinking, and rehearse how they will enact the assessment in class.• Consider feedback strategies might be appropriate given the anticipated student responses to the formative assessments
9	4/24/12	Reflect	<ul style="list-style-type: none">• Discuss enactment of first formative assessment activity• Complete revisions around video formative assessment
10	5/29/12	Reflect	<ul style="list-style-type: none">• Debrief formative assessment activity• Plan for next year
11	9/5/12	Explore student ideas	<ul style="list-style-type: none">• Plan for the upcoming school year• Generate a well-articulated explanation of natural selection and compare to learning progression.
12	10/3/12	Reflect and revise	<ul style="list-style-type: none">• Initiate the "reflect and revise" part of the professional development cycle• Revisit the formative assessments from last year and student responses to them

(Continued)

APPENDIX: CONTINUED

Session	Date	Phase	Activities
13	11/7/12	Reflect and revise, develop tools	<ul style="list-style-type: none"> ■ Introduce new learning progression document ■ Watch and reflect upon video clips of students doing the formative assessments from last year ● Identify different student ideas in the videotapes of students working on the formative assessments ● Connect what we saw in the video with the talk moves from <i>Ready, Set, Science!</i> (Michaels, Shouse, & Schweingruber, 2007) ● Discuss implications for the formative assessment design ● Suggest revisions for formative assessments based upon <i>Ready, Set, Science!</i> and what we saw in the video
14	12/5/12	Reflect and revise, develop tools	<ul style="list-style-type: none"> ● Reflect upon how the trial enactment of the formative assessment approaches went ● Focus on student products of formative assessment, the practices around the assessment ● Consider “what we learned” ■ Talk about instructional next steps
15	12/10/12	Reflect and revise, develop tools	<ul style="list-style-type: none"> ● Provide teachers with assessment reports for their 2012 students and explore them with reference to the learning progression document to focus teachers on: <ul style="list-style-type: none"> ○ What their students know this year based on the pretest results ○ What they can do instructionally to help students advance in their learning ● Review the purpose of formative assessment and springboard into the redesign for instructional experiences and formative assessment for spring 2013 ● Discuss formative assessment and instructional sequencing for the evolution unit with a focus on noticing and attending to student ideas
16	1/16/13	Reflect and revise, develop tools	<ul style="list-style-type: none"> ● Reflect back upon the different conversations, tools, and resources that we have used/built at the meetings this year. ● Plan the evolution unit conceptually to help students build a well-articulated explanation of natural selection and to leave days to build in feedback to assess student understanding during the unit. ● Plan when (and, if time, how) to do formative assessment and what concepts it will address in the unit.

(Continued)

APPENDIX: CONTINUED

Session	Date	Phase	Activities
17	2/6/13	Develop tools, practice using tools	<ul style="list-style-type: none"> Continue to plan formative assessments and the natural selection unit Plan formative assessment activities and their requisite anticipated feedback for the unit
18	3/6/13	Develop tools, practice using tools	<ul style="list-style-type: none"> Discuss dates the evolution unit will start Deeply read and interpret the learning progression and assessment report to determine what they suggest we do in planning formative assessments and the unit Draft formative assessments, feedback activities, and sketch in the activities that will go around them in the evolution unit
19	4/3/13	Develop tools, practice using tools	<ul style="list-style-type: none"> Review and revise the set of formative assessments we will use in the natural selection unit this year Develop a data analysis plan that will serve as a guide for interpreting data about student understanding collected on the formative assessments Produce two to three formative assessments with matched data analysis plans
20	5/1/13	Reflect on enactment	<ul style="list-style-type: none"> Engage teachers in data about what students know based upon what was revealed in the formative assessment that just happened Focus on student thinking, comparing evidence of student thinking to the learning progression, and using student thinking as resources for providing feedback and planning instruction.
21	5/29/13	Reflect on enactment	<ul style="list-style-type: none"> Explore posttest results Discuss relationship of posttest results to enactments of formative assessment during unit

We thank Michael J. Ross for his insights and thoughtful contributions to the development of the Elevate learning progression, the DANS, and an earlier draft of this paper, as well as the anonymous reviewers, whose thoughtful comments greatly improved the quality and clarity of the manuscript.

REFERENCES

- Alonzo, A. C., & Steedle, J. T. (2009). Developing and assessing a force and motion learning progression. *Science Education*, 93(3), 389–421.
- Anderson, D. L., Fisher, K. M., & Norman, G. J. (2002). Development and evaluation of the conceptual inventory of natural selection. *Journal of Research in Science Teaching*, 39(10), 952–978.
- Bennett, R. E. (2011). Formative assessment: a critical review. *Assessment in Education: Principles, Policy & Practice*, 18(1), 5–25.
- Bishop, B. A., & Anderson, C. W. (1990). Student conceptions of natural selection and its role in evolution. *Journal of Research in Science Teaching*, 27(5), 415–427.

- Board on Science Education. (2012). A framework for K-12 science education: Practices, crosscutting concepts, and core ideas. Washington, DC: National Academies Press.
- Borko, H., Jacobs, J., Eiteljorg, E., & Pittman, M. (2008). Video as a tool for fostering productive discussions in mathematics professional development. *Teaching and Teacher Education*, 24(2), 417–436.
- Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment*, 11(1), 33–63.
- Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice*, 5–12.
- Catley, K., Lehrer, R., & Reiser, B. (2005). Tracing a prospective learning progression for developing understanding of evolution. Paper commissioned by the National Academies Committee on Test Design for K-12 Science Achievement.
- Coffey, J. E., Hammer, D., Levin, D. M., & Grant, T. (2011). The missing disciplinary substance of formative assessment. *Journal of Research in Science Teaching*, 48(10), 1109–1136.
- Corcoran, T., Mosher, F. A., & Rogat, A. (2009). Learning progressions in science: An evidence-based approach to reform. Philadelphia: Consortium for Policy Research in Education.
- Dagher, Z. R., & Boujaoude, S. (2005). Students' perceptions of the nature of evolutionary theory. *Science Education*, 89, 378–391.
- Dobzhansky, T. (1973). Nothing in biology makes sense except in the light of evolution. *The American Biology Teacher*, 35, 125–129.
- Draney, K. (2009). Designing learning progressions with the BEAR assessment system. Paper presented at Learning Progressions in Science Conference, Iowa City, IA.
- Duncan, R. G., & Hmelo-Silver, C. E. (2009). Learning progressions: Aligning curriculum, instruction, and assessment. *Journal of Research in Science Teaching*, 46(6), 606–609.
- Duncan, R. G., Rogat, A. D., & Yarden, A. (2009). A learning progression for deepening students' understandings of modern genetics across the 5th–10th grades. *Journal of Research in Science Teaching*, 46(6), 655–674.
- Duschl, R., Maeng, S., & Sezen, A. (2011). Learning progressions and teaching sequences: A review and analysis. *Studies in Science Education*, 47(2), 123–182.
- Evans, E. M., Spiegel, A. N., Gram, W., Frazier, B. N., Tare, M., Thompson, S., & Diamond, J. (2010). A conceptual guide to natural history museum visitors' understanding of evolution. *Journal of Research in Science Teaching*, 47(3), 326–353.
- Ferrari, M., & Chi, M. T. H. (1998). The nature of naive explanations of natural selection. *International Journal of Science Education*, 20(10), 1231–1256.
- Furtak, E. M. (2009). Toward learning progressions as teacher development tools. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science conference*. Retrieved April 15, 2013, from <http://education.msu.edu/projects/leaps/proceedings/Default.html>.
- Furtak, E. M. (2012). Linking a learning progression for natural selection to teachers' enactment of formative assessment. *Journal of Research in Science Teaching*, 49(9), 1181–1210.
- Furtak, E. M., Morrison, D. M., Iverson, H., Ross, M., & Heredia, S. C. (2011). A conceptual analysis of the conceptual inventory of natural selection: Improving diagnostic utility through within-item analysis. Paper presented at the annual meeting of the National Association of Research in Science Teaching Annual Meeting, Orlando, FL.
- Garvin-Doxas, K., & Klymkowsky, M. (2008). Understanding randomness and its impact on student learning: Lessons learned from building the Biology Concept Inventory (BCI). *Life Sciences Education*, 7(2), 227–233.
- Geraedts, C. L., & Boersma, K. T. (2006). Reinventing Natural Selection. *International Journal of Science Education*, 28(8), 843–870.
- Grant, P. (1986). *Ecology and evolution of Darwin's finches*. Princeton, NJ: Princeton University Press.
- Gunckel, K. (2013). Teacher knowledge for using learning progressions in classroom instruction and assessment. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.
- Gunckel, K. L., Covitt, B. A., Salinas, I., & Anderson, C. W. (2012). A learning progression for water in socio-ecological systems. *Journal of Research in Science Teaching*, 49(7), 843–868.
- Heredia, S., Furtak, E. M., & Morrison, D. (2012). Item context: How organisms used to frame natural selection items influence student responses choices. Paper presented at the annual meeting of the National Association of Research in Science Teaching, Indianapolis, IN.
- Heritage, M. (2008). *Learning progressions: Supporting instruction and formative assessment*. Washington, DC: Council of Chief School Officers.
- Heritage, M., Kim, J., Vendlinski, T., & Herman, J. (2009). From evidence to action: A seamless process in formative assessment? *Educational Measurement: Issues and Practice*, 28(3), 24–31.
- Jin, H., Choi, J., & Anderson, C. W. (2009). Development and validation of assessments for a learning progression on carbon cycling in socio-ecological systems. Paper presented at Learning Progressions in Science Conference.

- Johnson, P., & Tymms, P. (2011). The emergence of a learning progression in middle school chemistry. *Journal of Research in Science Teaching*, 48(8), 849–877.
- Lehrer, R., & Schauble, L. (2012). Seeding evolutionary thinking by engaging children in modeling its foundations. *Science Education*, 96(4), 701–724.
- Mayr, E. (1982). *The growth of biological thought: Diversity, evolution, and inheritance*. Cambridge, MA: The Belknap Press of Harvard University Press.
- Merriam, S. B. (1998). *Qualitative research and case study applications in education*. San Francisco: Jossey-Bass.
- Metz, K. E., Sisk-Hilton, S., Berson, E., & Ly, U. (2010). Scaffolding children's understanding of the fit between organisms and their environment in the context of the practices of science. In K. Gomez, L. Lyons, & J. Radinsky (Eds.), *Learning in the disciplines: Proceedings of the 9th International Conference of the Learning Sciences (ICLS 2010)*, Volume 1, Full Papers. Chicago: International Society of the Learning Sciences.
- Michaels, S., Shouse, A. W., & Schweingruber, H. A. (2007). *Ready, set, SCIENCE!* Washington, DC: National Academies Press.
- Minstrell, J., Anderson, R., Kraus, P., & Minstrell, J. (2008). From practice to research and back: Perspectives and tools in assessing for learning. In J. Coffey, R. Douglas, & C. Stearns (Eds.), *Assessing science learning: Perspectives from research and practice* (pp. 37–68). Arlington, VA: National Science Teachers' Association Press.
- Mohan, L., & Anderson, C. W. (2009). Teaching experiments and the carbon cycle learning progression. Paper presented at Learning Progressions in Science Conference.
- Mohan, L., Chen, J., & Anderson, C. W. (2009). Developing a multi-year learning progression for carbon cycling in socio-ecological systems. *Journal of Research in Science Teaching*, 46(6), 675–698.
- National Research Council. (1996). *The National Science Education Standards*. Washington, DC: National Academy Press.
- National Research Council. (2007). *Taking science to school: Learning and teaching science in Grades K-8*. Washington, DC: National Academies Press.
- Nehm, R. H., & Ha, M. (2011). Item feature effects in evolution assessment. *Journal of Research in Science Teaching*, 48(3), 237–256.
- Odom, A. L., & Barrow, L. H. (1995). Development and application of a two-tier diagnostic test measuring college biology students' understanding of diffusion and osmosis after a course of instruction. *Journal of Research in Science Teaching*, 32(1), 45–61.
- Otero, V. K., & Nathan, M. J. (2008). Preservice elementary teachers' views of their students' prior knowledge of science. *Journal of Research in Science Teaching*, 45(4), 497–523.
- Rudolph, J. L., & Stewart, J. (1998). Evolution and the nature of science: On the historical discord and its implications for education. *Journal of Research in Science Teaching*, 35(10), 1069–1089.
- Sadler, P. M. (1998). Psychometric models of student conceptions in science; Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in Science Teaching*, 35(3), 265–296.
- Sherin, M. G., & Han, S. Y. (2004). Teacher learning in the context of a video club. *Education*, 20, 163–183.
- Shtulman, A. (2006). Qualitative differences between naive and scientific theories of evolution. *Cognitive Psychology*, 52, 170–194.
- Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic-molecular theory. *Measurement: Interdisciplinary Research & Perspective*, 4(1–2), 1–98.
- Songer, N. B., & Gotwals, A. W. (2012). Guiding explanation construction by children at the entry points of learning progressions. *Journal of Research in Science Teaching*, 49(2), 141–165.
- Songer, N. B., Kelcey, B., & Gotwals, A. W. (2009). How and when does complex reasoning occur? Empirically driven development of a learning progression focused on complex reasoning about biodiversity. *Journal of Research in Science Teaching*, 46(6), 610–633.
- Steedle, J. T., & Shavelson, R. J. (2009). Supporting valid interpretations of learning progression level diagnoses. *Journal of Research in Science Teaching*, 46(6), 699–715.
- Stevens, S. Y., Delgado, C., & Krajcik, J. S. (2010). Developing a hypothetical multi-dimensional learning progression for the nature of matter. *Journal of Research in Science Teaching*, 47(6), 687–715.
- Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.
- Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching*, 46(6), 716–730.
- Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education*, 13(2), 181–208.
- Wiser, M., Smith, C. L., Doubler, S., & Asbell-Clarke, J. (2012). Learning progressions as tools for curriculum development: Lessons from the inquiry project. Paper presented at the Learning Progressions in Science (LeaPS) Conference, June 2009, Iowa City, IA.
- Yin, R. K. (2003). *Case study research: Design and methods* (3rd ed.). Thousand Oaks, CA: Sage.

Creating Opportunities for Students to Show What They Know: The Role of Scaffolding in Assessment Tasks

HOSUN KANG,¹ JESSICA THOMPSON,² MARK WINDSCHITL²

¹*School of Education, University of California–Irvine, Irvine, CA 92697, USA;* ²*College of Education, University of Washington, Seattle, WA 98195, USA*

Received 26 March 2013; accepted 27 March 2014

DOI 10.1002/sce.21123

Published online 22 May 2014 in Wiley Online Library (wileyonlinelibrary.com).

ABSTRACT: This study examines the ways in which teachers provide students with written scaffolds in assessment tasks and the impact of these on students' abilities to demonstrate a core disciplinary proficiency—constructing evidence-based explanations. Data include 76 assessment tasks designed by 33 science teachers and 707 samples of student work. We found five types of scaffolding embedded in assessments that allowed students to make their reasoning explicit: (a) using contextualized phenomena, (b) rubrics, (c) checklists, (d) sentence frames, and (e) encouraging students to draw explanatory models in combination with written explanation. Analyses showed that all five forms of scaffolding were significantly associated with the quality of student explanation even when controlling for teacher variance and student background. Providing contextualized phenomena had the greatest impact on the quality of student explanations, both by itself and in combination with other scaffolding. The results indicate that strategic combinations of scaffolds can prompt students across all achievement levels to more readily use what they know to produce evidence-based explanations, but that the scaffolding must be of high quality.

© 2014 Wiley Periodicals, Inc. *Sci Ed* 98:674–704, 2014

Correspondence to: Hosun Kang; e-mail: hosunk@uci.edu

Contract grant sponsor: National Science Foundation.

Contract grant number: DRL-0822016.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of these organizations.

INTRODUCTION

Reform visions of science learning that focus on explanatory reasoning about complex bodies of knowledge raise expectations for student performance beyond basic competency levels (National Research Council [NRC], 2012a; Resnick, 2010). These expectations require new forms of teaching expertise in which professionals frame challenging tasks for students while creating varied opportunities for them to demonstrate what they know. Teachers can support both students' intellectual engagement and their demonstration of deeper learning by providing well-designed assessments.

Yet the power of assessment to reveal and support learning depends on how well student responses to tasks authentically reflect their thinking and understanding (Shepard, 2005). When such assessments are well designed, teachers gain insights into students' current ideas, gaps in understanding, and reasoning processes. With this information, teachers can adapt instruction based on learners' needs and strategically move students toward more advanced thinking. We refer here to structured forms of written assessment (as opposed to assessing through instructional conversations, e.g.) that are designed in advance by the teacher (see Ruiz-Primo & Furtak, 2007). Depending upon their scope and timing, such assessments may be considered formative or summative. However for the purposes of this study we make the assumption that revealing what students know is important at multiple times throughout a unit. In many of the cases we present later, the teacher has administered the assessment near the end of a unit (normally a characteristic of summative assessment) but used the responses to understand the impact of their instructional choices as well as to provide valuable feedback to students on their thinking (normally a characteristic of formative assessment). In this study then, we will focus more on *the potential of tasks to illuminate students' scientific understandings and abilities to construct evidence-based explanations*, and less on the question about where the use of such tasks might be labeled as formative or summative.

Little is known about the types of scaffolding teachers use within assessments, especially when designing written tasks, and how this can support student learning. We do know that the design qualities of the assessments themselves significantly affect the quality of produced responses (Herman, 1992; Supovitz, 2012), especially when the task requires higher levels of intellectual work and language use. Most previous research into the design qualities of assessments, however, has focused on externally developed assessments, not tasks designed by teachers (Supovitz, 2012).

In this study, we examined "medium-cycle assessments" (Supovitz, 2012) that teachers construct and administer within a unit of instruction to provide information about students' understanding of content. All our participating teachers used a framework to organize sequences of learning activities that were intended to support students' understanding of "big" science ideas by engaging them in the construction and revision of evidence-based explanations for a selected phenomenon throughout the unit (Windschitl, Thompson, Braaten, & Stroupe, 2012). The assessment tasks analyzed in this study were given to students typically 1 or 2 days before the last day of the unit to make the progress of students' ideas visible. The teachers used the information from the assessments in varied and productive ways, such as providing specific feedback to students on their responses, reviewing previous activities that students were confused about, having students talk to peers who constructed different explanations, and encouraging students to revise their explanation. We focused on teachers' use of scaffolding in assessment tasks to create opportunities for students to show what they had learned through 2 or 3 weeks of instruction. Specifically, we explore the impact on students' construction of evidence-based explanations—a core proficiency as described in recent consensus reports (NRC, 2012b). We examined the scaffolding embedded in 76

assessment tasks designed by science teachers and 707 samples of student work. These samples represent a range of student academic backgrounds (i.e., students who learn with difficulty, who are typical, and who learn easily).

THEORETICAL FRAMEWORK

Scaffolding and Assessment Tasks

Contemporary learning theories highlight two key ideas about the process of learning. First, learning involves the construction of knowledge. Second, learning and development are culturally embedded, socially supported processes (Bransford, Brown, & Cocking, 1999; Sawyer, 2006; Shepard, 2005). When learners actively participate in constructing knowledge in a culturally and socially supported learning environment, they gain deeper understandings, more generalizable knowledge, and greater motivation to use that knowledge in other settings (Brown & Campione, 1994; Smith, Maclin, Houghton, & Hennessey, 2000). These insights into the processes of learning draw researchers' attention to essential supports, or scaffolding, which must be provided in the learning environment.

Drawing upon sociocultural learning theory and Vygotsky's (1978) zone of proximal development, Shepard (2005) points out that scaffolding and formative assessment are "essentially the same thing" (p. 66) in that both are strategies that teachers use to move learning forward within the zone of proximal development. In formative assessment, teachers collect and analyze information about students' learning, then use that information to provide additional support to meet students' changing needs. With supports, learners can achieve a learning goal or produce what they cannot produce alone (or can only with difficulty). Scaffolding refers to various forms of material, social, linguistic, or conceptual assistance that can support students' reasoning, participation, and learning (Sawyer, 2006). When teachers engage in effective forms of assessment, they are likely to provide scaffolding, such as prompts or response structures that address learners' difficulties and are informed by previous student responses and classroom talk.

A critical component of scaffolding that distinguishes it from other forms of support is "fading" (Pea, 2004). According to Pea, work on tasks that are assisted by scaffolding should be achieved later without such assistance when the learner becomes proficient. From an instructional point of view, the decision of when and how to remove a support is crucially important in shaping learning opportunities—as much as the decisions about the forms of scaffolding themselves. In a science classroom, fading will be dependent upon the complexity of the tasks as well as the degree of learners' progress. Constructing evidence-based explanation is not only complex and intellectually challenging but also new to most students in K-12 science classrooms. The literature is unclear about forms of scaffolding that assist students' mastery of such intellectually challenging tasks, and how the scaffolds might be faded.

Studies of English language learners (ELLs) provide particularly useful insights of scaffolding in science classrooms because engaging with wide-ranging bodies of knowledge and the explanatory practices of the discipline impose significant linguistic demands on these students. The theories behind ELL scaffolding, however, are applicable to all students. Walqui (2006) proposed six types of instructional scaffolding for ELLs. Four of the six are particularly relevant to constructing evidence-based explanations. First, *instructional modeling* is providing clear examples of what is requested of students for emulation. The objects of this modeling (not to be confused with scientific modeling) encompass tasks and activities, but also "appropriate language use for the performance of specific academic functions" (p. 171). Constructing evidence-based explanations requires particular ways of

using language, and instructional modeling as one form of scaffolding enables teachers to help students employ appropriate disciplinary discourses. *Bridging* refers to providing support that helps students connect their previous knowledge and understandings with new concepts and language. An important aspect of bridging is establishing a personal link between the student and the subject matter by “showing how new material is relevant to the student’s life, as an individual, here and now” (p. 172). Traditional science teaching frequently fails to make connections to children’s day-to-day lives (Moje et al., 2004; Warren, Ballenger, Ogonowski, Rosebery, & Hudicourt-Barnes, 2001). *Contextualizing* refers to the use of language in concrete sensory contexts. Examples of this form of scaffolding include using material manipulation, pictures, a few minutes of a film, and other types of realia. In science, students need help working with the decontextualized, situation-independent, and dense academic language; contextualizing makes ideas and language more accessible and engaging for students. Finally, *developing metacognition* involves supporting learners’ ability to monitor their current level of understanding and to decide when it is not adequate for a specified task. Examples of scaffolding here are rubrics or lists of steps for the routine being practiced. Variants and combinations of these forms of support can make it possible for students to productively engage in cognitively and linguistically challenging assessment tasks (Nasir, Rosebery, Warren, & Lee, 2006).

Disciplinary Proficiency Projected in a Written Assessment Work: Constructing Evidence-Based Explanations

The construction of explanations is an essential feature of science, as well as a fundamental classroom activity that engages students in epistemic practices of the discipline (Knorr-Cetina, 1999; Latour & Woolgar, 1979; Nersessian, 2005; Pickering, 1995). Recent reform documents, including the *Next Generation Science Standards* and its associated *K-12 Framework*, have highlighted causal explanation as a central practice (NRC, 2012b, 2013). There is a growing consensus among science educators that student-produced explanations—as opposed to the reproduction of textbook explanations—are evidence of deep science learning (Braaten & Windschitl, 2011; Ford & Wargo, 2012; NRC, 2000, 2007, 2012b). Written explanations, as artifacts of this core disciplinary practice, reflect not only students’ conceptual understanding and reasoning but also their epistemic commitment to a specialized form of knowledge building (Ford & Wargo, 2012; Sandoval, 2003).

Scientific explanations are causal accounts for phenomena reflecting how one makes sense of events and processes in the natural world. Thoughtful scientific explanations articulate the underlying mechanisms, going beyond “explicating” observable phenomena to “theorizing” how and why things happen (Braaten & Windschitl, 2011). Constructing scientific explanation involves positing particular kinds of relationships, specifically that natural processes or events are attributed to a set of factors that produced the phenomenon (Ohlsson, 2002). During the process, unobservable mechanisms are distinguished from observable events or processes and connected to each other through coherent and principled reasoning. When students (and scientists) construct scientific explanations, they can engage in either reasoning from data or with known scientific ideas to develop/revise a coherent, causal explanatory model (“explanatory hypothesis”) about how and why things happen. In this way, justifying elements of an explanatory model with sufficient evidence is one essential feature of scientific explanation. A good scientific explanation accounts for patterns in data and links claims in these accounts with relevant evidence (Sandoval & Millwood, 2005; Sandoval & Reiser, 2004).

In this study, we use the idea of “evidence-based explanation” rather than just “scientific explanation” to identify core disciplinary proficiencies projected in a written assessment

task. Currently, the meaning of scientific explanation is underconceptualized, especially in relation to the practice of argumentation with evidentiary warrants.¹ Many researchers have combined the goals of explanation and argumentation, and then characterized scientific explanation as that which justifies explanations of scientific phenomena where claims are supported with appropriate evidence and reasoning (Furtak & Ruiz-Primo, 2008; McNeil & Krajcik, 2006; Ruiz-Primo, Li, Tsai, & Schneider, 2010). The quality of student-written explanations, then, have been examined referencing three components—claim, evidence, and reasoning—that originated from Toulmin’s (1958) argument structure (Furtak & Ruiz-Primo, 2008; Ruiz-Primo et al., 2010). For example, Ruiz-Primo and colleagues (2010) evaluated the quality of student explanations in terms of (a) focus and accuracy of the claim, (b) type, nature, and sufficiency of evidence, and (c) the alignment and quality of the link between evidence and claim in reasoning. Although the three components are frequently cited as essential in scientific explanations (Kenyon & Reiser, 2006; McNeil & Krajcik, 2006; Sandoval & Reiser, 2004), using the components of argumentation to evaluate the quality of scientific explanations gives rise to several issues, such as looking past students’ opportunities to theorize “how and why” things happen—beyond justifying a claim.

By using the idea of evidence-based explanation, we intend to examine the quality of student explanations, considering students’ capability to theorize how and why things happen as well as justify their working theories with the use of evidence. The following describes these criteria in detail.

Conceptualizing the Quality of Student Explanations

Building on previous studies and grounded in the analysis of 707 samples of student work, we propose four dimensions of explanation that mark important differences in the quality of written accounts:

1. *The conjectural framing of explanations*: How observable natural phenomena and unobservable processes or ideas are treated in the explanation
2. *The role of evidence*: How explanations are supported with observation or data
3. *The depth of explanation*: The degree to which explanations provide comprehensive and gapless accounts for focal phenomena, including causal relationships and underlying mechanisms
4. *Causal coherence*: How explanations are logically consistent with data, observation, evidence (i.e., internal coherence) as well as generally accepted scientific principles and theories (i.e., external consistency)

¹Scientific explanation is similar to argumentation in that both practices involve reasoning from the data, giving special credit to evidentiary support and generating a tentative conclusion. In the case of scientific explanation, this tentative conclusion is perceived as a current best explanatory hypothesis. However, scientific explanation and argumentation differ in their goals and the entities that invoke each practice. The goal of scientific explanation is “To provide an account that offers a plausible causal mechanism” about a natural phenomenon (Osborne & Patterson, 2011, p. 634). The entity for scientific explanation is the feature or phenomenon that is observed. Therefore, typical scientific explanations consist of a statement of the feature or a phenomenon and causal accounts for it. In contrast, the goal of argumentation is to persuade or “To provide incontrovertible warrants that support the claim and to show that it is a justified belief” (Osborne & Patterson, 2011, p. 634). The entity that invokes argumentation is the validity of claims or any explanation, not natural events or phenomena. The primary focus of argumentation is to examine the link between claims and evidentiary warrants (i.e., whether the claim is well supported with quality evidence) with the goal of evaluating or justifying the validity of claim, not understanding the relevant natural phenomena.

This framework guides the evaluation of student explanations. The following describes four dimensions of evidence-based explanations that reflect the quality of reasoning by students.

The Conjectural Framing of Explanation. Explanations can be characterized as either narrated or constructed as reflecting the modes of thought involved in the processes (Bruner, 1985). *Narrated explanations* take a form of a “correct version” of a story about some natural phenomenon. Students (re)produce uniform textbook-like explanations without significant variation. The links between observable phenomena and unobservable scientific ideas are not clear, and the tentative, revisable, and testable features of the explanation are not evident. In contrast, *constructed explanations* show causal links between observable phenomena and a proposed explanatory mechanisms. These explanations incorporate claims and reasoning. In some cases, students’ explanations are constructed by reasoning through data and at other times by principled reasoning with scientific ideas. Students usually reveal a wider spectrum of understandings when they construct explanations about natural phenomena as opposed to reproducing textbook explanations.

The Role of Evidence. The second dimension of explanation is the role that evidence plays. Evidentiary support is a key feature of explanation that indicates students’ conceptions of the epistemic nature of the discipline (McNeil & Krajcik, 2006; Sandoval, 2003). In science explanation studies, two characteristics of evidence—appropriateness and sufficiency—are frequently used (Kenyon & Reiser, 2006; McNeil & Krajcik, 2008; Sandoval, 2003). Appropriateness concerns whether the data cited are relevant to the problem, and sufficiency involves whether sufficient and credible data are provided to warrant the claim. Ruiz-Primo and her colleagues extended this framework to better capture the quality of evidence. They examined (a) type (i.e., what type of evidence did the student provide?), (b) nature (i.e., did the student focus on patterns of data or isolated examples?), and (c) sufficiency (i.e., did the student provide enough evidence to support the claim?). Building on these previous studies, we have characterized explanations with special attention to the roles of evidence: (a) explanations with no support (i.e., no use of evidence), (b) evidence-referring explanations, and (c) evidence-based explanations. *Evidence-referring explanations* have some forms of evidence, but the connection between evidence and claim is not sufficient or appropriate. For example, students simply refer to activities, information, or data as evidence without elaborating key patterns of the activity and how they support their claim. In contrast, *evidence-based explanations* are supported by data or observations that are directly related to the object of explanation.

The Depth of Explanations. The third dimension of student explanation is the degree to which explanations provide comprehensive and in-depth accounts of observable phenomena or events. The depth of explanations—providing a best explanatory model of how and why things happen—is the heart of scientific explanation. Drawing upon Braaten and Windschitl’s (2011) framework, we have characterized explanations as (a) “what” explanations, (b) simple causal explanations or “how” explanations, and (c) “why” explanations. “What” explanations focus on describing observations in terms of patterns, without suggesting cause. *Simple causal explanations* focus on a causal relationship of one observable event affecting another with little attention to underlying mechanisms or principles (i.e., what fundamentally influences observations). “Why” explanations provide causal stories for a phenomenon and use unobservable or theoretical ideas or processes to explain patterns of observations.

Causal Coherence. A good scientific explanation is internally consistent, thus providing well-connected accounts for focal phenomena. Such an explanation is also consistent with generally accepted scientific principles and theories. Using causal coherence as an analytic feature, student explanations can be characterized as (a) explanations with no causation expressed, (b) partially coherent explanations, or (c) coherent and consistent causal explanations. First, *explanations with no causation* have bits and pieces of information relevant to the problem, but the information is not organized so as to reveal the causal relationships. Such an explicating-type explanation (i.e., “what” explanation) is likely to provide no causation. *Partially coherent explanations* show a chain of reasoning that is not fully coherent internally or not consistent with other science ideas. For example, students may produce an interesting and plausible theory of how and why things happen, but such theory may be scientifically inaccurate (i.e., inconsistent with generally accepted scientific ideas). In other cases, some part of the explanation may be coherent and logical, but other parts conflict with the referenced data or evidence. *Coherent and consistent causal explanations* show a logical chain of reasoning; the link between evidence and an explanatory model is adequate and substantial (i.e., internally consistent), and the proposed explanatory model is consistent with generally accepted scientific ideas (i.e., externally consistent).

The Use of Scaffolding and Opportunities to Construct Evidence-Based Explanations

Studies have documented various forms of scaffolding used to support students’ authentic disciplinary talk and writing. Overall, the scaffolding discussed in the science education literature can be characterized as structure-oriented claim, evidence, reasoning (CER) and explanation-oriented scaffolding.

Structure-oriented CER scaffolding is the most popular form of support and focuses primarily on providing the *structure* of scientific explanations inspired by Toulmin. Sometimes, researchers also provide additional support to help students incorporate concepts into explanations (see Kenyon & Reiser, 2006; McNeil & Krajcik, 2006; Songer & Gotwals, 2012). For example, Songer and Gotwals (2012) investigated how “scaffold-rich assessments” support young students’ explanations. One task was to construct an explanation given a specified scientific question, “Is the large fish [in an ecosystem] a producer or a consumer?” This assessment task was laid out with three boxed spaces that had the subheadings of “Make a CLAIM,” “Give your REASONING,” “Give your EVIDENCE.” The authors also provided detailed prompts for each component (e.g., “Make a CLAIM: Write a sentence that answers the scientific questions”) along with concept support (e.g., “Hint: Think about how producers and consumers get energy”). Similarly, Kenyon and Reiser (2006) provided both an “explanation framework” consisting of CER and criteria for evaluating the quality of each structural component. CER scaffolding provides opportunities for students to structure their explanations, reminding them of relevant scientific concepts that they need to consider. Students are provided opportunities to learn what scientific explanations look like and to craft their explanations following this guidance. Using these supports, Songer and Gotwals (2012) reported that students’ conceptual understanding increased based on pre- and posttest comparisons.

Explanation-oriented scaffolding has a similar objective to CER structure-oriented scaffolding but tends to feature the blending of conceptual and epistemic scaffolds. More of an emphasis is placed on explaining authentic scientific phenomena, rather than a single lab-based classroom experiences. For example, Sandoval and Reiser (2003, 2004) developed a computer-based tool, “Explanation Constructor,” to support students’ construction and evaluation of explanations about a situated example of natural selection (e.g., “Why are the

finches that survive able to survive?”). Explanation Constructor provides conceptual scaffolds, such as a series of sentence frames that guide students to consider key patterns (e.g., “The existing variation in the population before the pressure is . . .”), how things happen (e.g., “How has the distribution of organisms in the population with this trait changed?”), and why (e.g., “The survivors are the most fit under this pressure because they have these traits . . . that enable them to . . .”). Explanation Constructor also provides evidence such as data or figures adjacent to the explanation space, thus prompting students to use evidence in writing their explanation. As Sandoval and Reiser (2004) noted, these computer-based scaffolds mediate students’ activity by acting as “enablers,” but their role depends on students’ understanding of the purpose of their work and the affordances for action that are shaped by social interaction between teacher and students.

The nature of opportunities created with the specific use of scaffolding has significant implications for the participation of students in disciplinary activities, especially students for whom science represents a different way of knowing, talking, or doing than is prevalent in their life experiences (Moje, Collazo, Carrillo, & Marx, 2001; Rosebery, Ogonowski, DiSchino, & Warren, 2010). The construction of written explanations involves negotiating meaning, mediated by new language and varied texts. Determining how to negotiate the multiple texts, discourses, and knowledge available within the learning community can be challenging to students from nondominant cultural and linguistic backgrounds (Moje et al., 2001; Rosebery et al., 2010). Importantly, well-designed scaffolding makes it possible to provide academically challenging instruction for students who typically are underserved in secondary schools (Walqui, 2006).

RESEARCH QUESTIONS

The following research questions are addressed:

1. What types and combinations of scaffolding do teachers most often use when designing written assessment tasks?
2. How does each type of scaffolding relate to the extent and quality of students’ scientific explanations?
3. Do certain levels of quality and combinations of scaffolding influence the quality of students’ explanations more than others?

METHODS

Research Context and Participants

We employed a mixed-methods approach (Creswell & Plano Clark, 2011) to study how and why particular forms of scaffolding embedded in assessments support students’ construction of written evidence-based explanations. Assessment tasks and samples of student work were collected from 33 first-year science teachers who participated in induction activities between 2010 and 2012. All the teachers graduated from a teacher education program at a public university in the United States between 2010 and 2011 and taught science at the secondary level in local communities. They received support from the university in their first year of teaching. Throughout the 2 years from preparation through the first year of teaching, they were exposed to the resources and tools for reform-oriented science teaching that emphasized students’ construction of evidence-based explanations. The teachers participated in three sessions of collegial analyses of student work artifacts, facilitated by the university research team during their first year of teaching (for details, see Thompson et al., 2009; Windschitl et al., 2011). On these occasions, teachers were asked to bring samples

TABLE 1
The Coding Scheme About the Characteristics of Scaffolding in Assessment Tasks

Type of Scaffolding	Level 0 (code = 0)	Level 1 (code = 1)	Level 2 (code = 2)
Drawing in combination with writing	No drawing	Generic drawing/ Posterizing	Modeling
Contextualizing phenomena	Generic phenomena	Contextualizing	NA
Checklist	No checklist	Simple words checklist	Explanation checklist
Rubric	No rubric	Generic rubric	Comprehensive rubric
Sentence frame	No sentence frame	Focusing	Connecting

of student explanations. Teachers were provided a rubric for evaluating samples of student work prior to the collaborative analysis of these artifacts. Specifically the rubric asked teachers to evaluate their own students’ work based on the (1) degree to which the student made comparisons among pieces of evidence, (2) degree of depth in student’s explanation, and (3) degree to which evidence and explanations were integrated in written products. In each session, teachers brought their assessment tasks along with 9–12 samples of work from students with varied academic backgrounds. They were asked to bring three to four samples of student work from students who appeared to learn new ideas easily, the same number from students who were in the midrange of academic achievement, and the same number who appeared to learn with difficulty in their classrooms. We also asked teachers to include one or two samples of students who had special needs, such as ELLs. The 76 assessment tasks were approximately evenly distributed across the 33 teachers. These assessments and the samples of student work from the induction activities became the sources of data.

Data Sources and Measures

We analyzed 76 assessment tasks and 707 copies of student work. The collected student work consisted of even percentages of students with different academic backgrounds (students who learn with difficulty (30.8%), typical students (29.3%), students who learn easily (29.7%)). Of the student work, 72 copies did not have identification (10.2%) and, thus, were excluded from regression analyses. The types of scaffolding in the assessment tasks and the quality of student explanation in student work were coded using schemes described in the following section—one set around scaffolding and one set around the quality of student explanation.

Use of Scaffolding in Written Assessment Tasks

We found five salient modes of scaffolding from this initial analysis: (a) allowing students to draw in combination with writing, (b) contextualizing the explanation within a focal phenomenon or event, (c) providing checklists, (d) using rubrics, and (e) providing sentence frames. Within each type of scaffolding, there existed different levels of design sophistication. The main features of each type of scaffolding, and how they were converted quantitatively, are shown in Table 1.

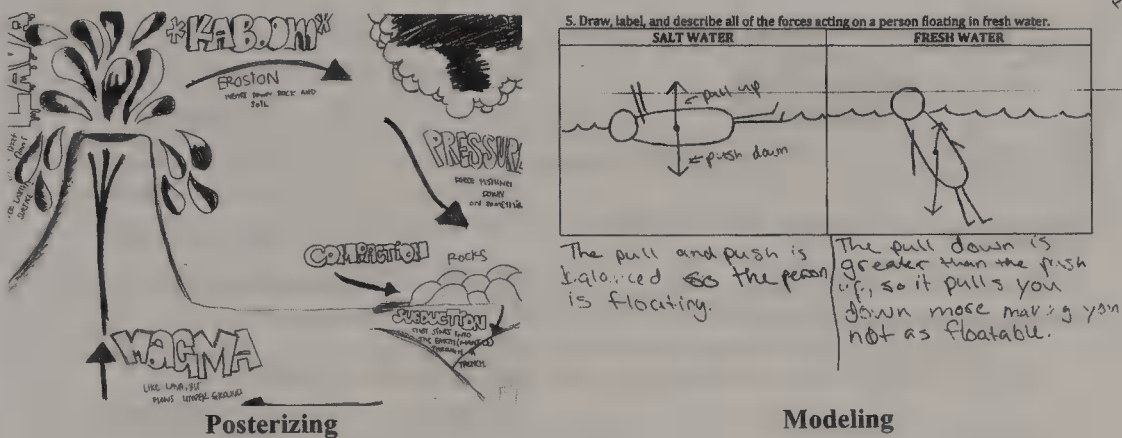


Figure 1. Drawing in combination with writing.

Scaffolding by Prompting Drawing in Combination With Writing. Many assessment tasks allowed students to explain focal phenomena with drawing and writing. We found two different categories for prompts for drawing: One was generic drawing or posterizing (these were at the lower level of sophistication), and the other was modeling (see both examples in Figure 1). Generic drawing asked students to illustrate any aspect of the focal phenomena without clear guidance. Posterizing prompted students to reproduce models that could be found in the textbook and that illustrated some known set of discrete/unproblematic relationships, such as in the rock cycle or a “standard” volcano eruption. Both generic drawing and posterizing were categorized as lower level (Level 1) because this form of drawing, as scaffolding, simply provided alternative ways to express canonical ideas. In contrast, when students engaged in modeling, the work of drawing itself engaged them in higher levels of intellectual work.

Drawing as modeling usually provided designated spaces, structures, or templates for drawing, such as boxes, an outline of the sun, human body, or an enlarged blank inset. In contrast, the samples of student drawing that were coded as posterizing often fail to include any structure or prompt in the task design as shown in Figure 1. A few noticeable characteristics appeared in the prompts for modeling. First, students were prompted to illustrate *unobservable underlying mechanisms* that cause an observable event or phenomenon. For example, in one assessment of cell membrane mechanisms in ninth-grade biology (see Figure 2), students were prompted to “draw [a] scientific diagram” showing “what is happening that we can’t see!” In some assessments, students were prompted to illustrate something over changes in time (e.g., draw what happens before, during, and after), temperature, and concentration (e.g., low vs. high). Another characteristic was that students were prompted to illustrate how events happened at an appropriate *scale* (e.g., cellular level, molecular level). Occasionally, students were not prompted to draw at a particular scale; in these cases, students did not use their drawing to explain ideas effectively. For example, in an assessment about cancer, students were prompted to draw what would happen to someone who had cancer before, during, and after, the onset of the disease but without specification of the scale. In this case, students illustrated a person at the organism level (i.e., drawing of human body), instead of what would happen at the cellular level, which made the drawings less useful for revealing in-depth explanations of cancer growth. We coded models using a scale of 0–2 (0 = no drawing, 1 = generic/posterizing, 2 = modeling) for regression analysis.

Scaffolding by Contextualizing a Focal Phenomenon. There was substantial difference in both the nature of explanation and the ways in which focal phenomena for explanation

Date: 4/10/16 Period: 9

How do Paramecium get everything they need to survive?

Part 1:
Paramecium live in pond water that is hypotonic to their single cell bodies. Draw a scientific diagram and write a full scientific explanation about how the Paramecium get the water and oxygen they need to survive. Also, explain what would happen to the Paramecium if all the water in their pond turned to salt water!

Use the answer checklist (on the back) and **idea checklist** (below) to help you. You may use your **idea checklist**. These ideas need to be included in your response. When using an idea, be sure to **explain what it means and why you are using it**.

<input checked="" type="checkbox"/> Diffusion	<input checked="" type="checkbox"/> Hypertonic	<input type="checkbox"/> Energy Required
<input type="checkbox"/> Osmosis	<input checked="" type="checkbox"/> Hypotonic	<input type="checkbox"/> Concentration Gradient
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/> Semipermeable Membrane

Scientific Diagram of Paramecium (Show what is happening that we can't see!)

Explanation
The paramecium is hypotonic, it had mean the water going into the paramecium, the water osmosis in the paramecium and the paramecium is bigger because the water molecules want from high to low and the water molecules don't take the energy from the water molecules moving in a high concentration solute. That is osmosis. The semipermeable membrane go through paramecium but some of semipermeable membrane go into the paramecium but some of semipermeable membrane can not go into the paramecium.

Answer Checklist: Be sure to check and make sure your explanation addresses and answers the following concepts.
1. Explain how paramecium get water to survive
2. Explain how paramecium get oxygen to survive
3. Explain what would happen to paramecium if salt water was added

Part 2: Using evidence to support your ideas
Once you have written your explanation, pick two pieces of evidence and write how they support your explanation.

Remember the evidence we have collected from each activity:

<input type="checkbox"/> Adding salt water to onion cells	<input type="checkbox"/> Break down and reassembly of food molecules over time
<input type="checkbox"/> The Hatched Egg in corn syrup, water, and <u>whites</u> .	<input type="checkbox"/> Transport of molecules through the cell <u>membrane</u> and <u>osmosis</u>
<input type="checkbox"/> Sugar-Water Osmosis Lab	
<input type="checkbox"/> Starch, Glucose, <u>osmosis</u> Lab	

Evidence for osmosis comes from the adding salt water to onion cells because the salt can be harmful to pull out the water inside the cell and can cause gaps at the cell.

Evidence for semipermeable membrane comes from the weight he came heavier because the water go into the sugar-water cell.

Figure 2. Cell membrane assessment: How does the paramecium get everything it needs to survive?

were framed in assessments. One group of assessment tasks asked students to explain general phenomena, “Why do siblings look different?,” “Why is the equator hotter than the poles?,” and “Why do the seasons change?” Often these assessments asked students to explain scientific ideas rather than an observable event, such as “What is homeostasis and why is it important to our body?” Representations of those events or phenomena usually appeared in the textbook. In contrast, some teachers contextualized a phenomenon or event in a particular time, place, and situation. For example, instead of asking about generic seasonal changes, one teacher asked, “Why don’t countries near the equator, like Samoa, seem to have seasons like we do here in Seattle?” In another assessment in a unit on force and motion for the seventh grade, a teacher contextualized the physics in the form of “the skater girl”—a young woman in a local community where the school is located (see Figure 3):

A skater girl is flying down the big hill on 102nd (right in front of Steve Cox Memorial Park, where that cabin is, behind McLendon’s Hardware) when she realizes that some jerk has built a huge brick wall across the road. She knows that she won’t be able to stop in time. What should she do to minimize, or decrease, her injuries? Explain why this is the best option for the skater girl.

For regression analysis, a noncontextualized idea or event in assessments was coded as 0 and the contextualized phenomenon or event was coded as 1.

Scaffolding by Providing a Checklist. The third form of scaffolding in the assessment task was providing one or a set of idea checklists to be referenced while constructing the explanation. We found two different kinds of checklists. The first is a “simple checklist” that lists concepts or scientific terms. This was often provided as a word bank in a box (see Figure 3). The other was an “explanation checklist” that prompted students to explain multiple aspects of the focal event as well as some relationships among ideas,

Name _____ Period 2 Date Feb 15th Score: ____/14

Final Explanation: The Skater Girl

THE SITUATION: A skater girl is flying down the big hill on 102nd (right in front of Steve Cox Memorial Park, where that cabin is, behind McLendon's Hardware) when she realizes that some jerk has built a huge brick wall across the road. She knows that she won't be able to stop in time. What should she do to minimize, or decrease, her injuries?

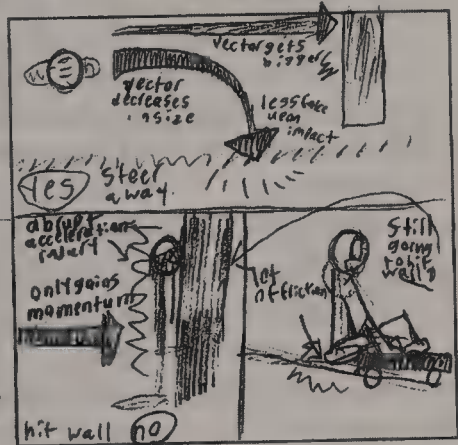
FINAL EXPLANATION: Use your journal, the Word Wall, and your Evidence Buckets to answer Questions 1–3.

1. What should the skater girl do to minimize her injuries? (1 point) steer towards the pond & grass for a softer landing than the wall
2. Explain, using words and pictures, why this is the best option for the skater girl. Use as many words from the word wall as you can. Use the space below to draw a picture that helps you answer it you need it. (3 points)

Word Bank			
Velocity	Vector	Acceleration	Momentum
Force	Net force	Friction	Mass

- 1 point: Describe the forces acting on the skater girl in #1.
- 1 point: Use at least 3 words from the Word Bank/Word Wall.
- 1 point: Explain why this is better than another choice she has.
- 1 point: Draw a picture that helps explain your answer.

The skater girl should try to (slowly) steer towards the grass, pond, or other surface to avoid the wall because steering (slowly) away would result in her accelerating in such a way that her velocity decreases & force upon impact isn't as great as it would be when hitting the wall. She shouldn't hit the wall, because the sudden change in momentum would hurt her, & she shouldn't drag her feet/butt because of the friction from the coarse cement would hurt her & she would still hit the wall.



USING EVIDENCE: Use your Evidence Buckets to answer this question.

3. Give at least one piece of evidence from a class activity that supports your ideas in #2. (2 points)
- The egg video thing shows that when momentum is changed very slowly, the stopping force upon said body is also very low. However, if momentum is changed abruptly, the stopping force is very high. Therefore, we know it would be a good idea for the girl to slowly change her momentum, so that her stopping force is small & injuries acquired are also fairly minor.

Figure 3. Force and motion assessment: What should the skater girl do to minimize her injuries?

observations, and key patterns. For example, in a cell membrane assessment (see Figure 2), a teacher asked students to explain how a paramecium survives in pond water, providing the following explanation checklist. This scaffolding was framed by the teacher as an “Answer Checklist”:

Answer Checklist: Be sure to check and make sure your explanation addresses and answers the following concepts:

- ☐ Explain how paramecium gets water to survive.
- ☐ Explain how paramecium gets oxygen to survive.
- ☐ Explain what would happen to paramecium if salt water is added.

The explanation checklist occasionally appeared along with the simple checklist. The checklist as a form of scaffolding was coded using a scale of 0–2: 0 = no checklist, 1 = simple checklist, and 2 = explanation checklist.

Scaffolding by Providing a Rubric. Rubrics provided information about the essential attributes of a high-quality evidence-based explanation. A “simple rubric” (Level 1) provided prompts, such as “Once you have written your explanation, pick two pieces of evidence and write how they support your explanation.” Sometimes, a Level 1 rubric provided a more elaborated list for getting full credit, as shown in the skater girl’s assessment (see Figure 3):

- 1 point: Describe the forces acting on the skater girl in #1.
- 1 point: Use at least three words from the Word Bank/Word Wall.
- 1 point: Explain why this is better than another choice she has.

A simple rubric is similar to checklist except that a rubric shows associated credit points. In contrast, a “comprehensive rubric” (Level 2) took the form of a table with multiple rows and columns, elaborating expected levels of performance in detail. The performance expectation was developed by the teacher and, thus, did not always match with our criteria for evaluating the quality of an explanation. We coded different forms of rubrics using a scale of 0–2: 0 = no rubric, 1 = simple rubric, and 2 = comprehensive rubric.

Scaffolding With Sentence Frames. Two different forms of sentence frames appeared in the assessments: focusing versus connecting. Focusing sentence frames (Level 1) prompted students to draw their attention to the phenomena and explain them by providing linguistic lead-ins. For example, in the assessment about the gas laws using the phenomenon of changes in a balloon sitting on a table, students were prompted with a sentence frame like this: “What I saw was _____,” “Inside [the balloon] the particles were _____,” and “I know this because _____.” In contrast to focusing sentence frames, connecting sentence frames (Level 2) prompted students to make deeper connections among key components of scientific explanation, such as evidence and reasoning. For example, in a cell membrane assessment (see Figure 2), the teacher provided the following prompt to support students’ use of evidence: Evidence for _____ comes from the _ [activity or reading] _ because ____.” Sentence frames were coded using a scale of 0–2: 0 = no sentence frame, 1 = focusing, and 2 = connecting.

Quality of Student Explanations

The quality of student responses was evaluated with respect to the four dimensions of good scientific explanation discussed in the conceptual framework (see Table 2). The four dimensions were (a) conjectural framing of explanation, (b) role of evidence, (c) depth of explanation, and (d) causal coherence. Each of the four dimensions was specified into three different levels. The first dimension, conjectural framing of explanation, had only two levels (0 and 1). For the three other dimensions, we assigned scores from 0–2 at each level depending on its sophistication (i.e., least sophisticated = 0, most sophisticated = 2). The unit of analysis was one sample of student work. The highest composite score was 7 when a sample was evaluated as being at the most sophisticated level of response across the four dimensions (i.e., a constructed, coherent and causal explanation that is strongly supported by evidence). The following shows three examples that illustrate different levels of sophistication (high, middle, and low). The first two examples are student explanation

TABLE 2
Criteria to Score Qualities of Explanations

Dimension	Level 1 (Code = 0)	Level 2 (Code = 1)	Level 3 (Code = 2)
Conjectural framing of explanation	<p>Narrated</p> <ul style="list-style-type: none"> Mostly restating textbook explanation as a form of an unproblematic story Observable phenomenon/event(s) is not distinguished from unobservable mechanism or scientific ideas 	<p>Constructed</p> <ul style="list-style-type: none"> Constructing explanation by reasoning through data Constructing explanation through principled reasoning (applying) using scientific theories/ideas/models 	
Role of evidence	<p>Explanation with no support</p> <ul style="list-style-type: none"> Explanation is not supported by any form of evidence 	<p>Evidence-referring explanation</p> <ul style="list-style-type: none"> Simply referring activities or data as evidence Some form of evidence is referred in the explanation, but the connection between evidence and explanation is weak. For example, referring the topic of activity rather than the key patterns of the activity that are related to the object of explanation 	<p>Evidence-based explanation</p> <ul style="list-style-type: none"> Highlighting key patterns of data or observations, or in activities to support a claim Explanation is supported by strong evidence that is directly related to the object of explanation
Depth of explanation	<p>“What” explanation</p> <ul style="list-style-type: none"> Describing observations in terms of patterns, without suggesting cause “Explanation as explication” (Braaten & Windschitl, 2011) 	<p>Simple causal explanation or “how” explanation</p> <ul style="list-style-type: none"> Focusing on a causal relationship of one observable event affecting another with little attention to underlying mechanisms or principles (i.e., what fundamentally influence observation) 	<p>“Why” explanation</p> <ul style="list-style-type: none"> In-depth constructed explanations that provide full causal stories for a phenomenon, and use unobservable or theoretical events or processes to explain patterns of observations

(Continued)

TABLE 2
Continued

Dimension	Level 1 (Code = 0)	Level 2 (Code = 1)	Level 3 (Code = 2)
		<ul style="list-style-type: none"> ▪ “Explanation as simple causation” (Braaten & Windschitl, 2011) 	
Causal coherence (logical consistency, internal, and external coherence)	Explanation with no causation <ul style="list-style-type: none"> ▪ Explanation covers bits or pieces of information about phenomena/events (illogical or incoherent list of information); there is no causal explanation about focal phenomena/event. 	Partially coherent explanation <ul style="list-style-type: none"> ▪ Partial explanation about the posed phenomena (“plausible-but-incorrect scientific explanation”); Some part of explanation is incoherent or illogical, conflict with data/evidence 	Coherent and consistent causal explanation <ul style="list-style-type: none"> ▪ Gapless and integrated explanation about the posed phenomena • Coherent and logical • No conflict with or among data/evidence (“causal coherence” internally consistent as well as consistent with generally accepted scientific principles and theories)

from the Skater Girl assessment (Figure 3), and the last one is a student response to an assessment task of a ninth-grade biology unit.

Example 1—highly sophisticated explanation: A constructed, evidence-based, and in-depth “why” explanation with coherent and integrated reasoning (score = 7 of 7)

The skater girl should try to (slowly) steer towards the grass, pond, or other surface to avoid the wall because steering (slowly) away would result in her accelerating in such a way that her velocity decreases & force upon impact isn’t as great as it would be when hitting the wall. She shouldn’t hit the wall, because the sudden change in momentum would hurt her, & she shouldn’t drag her feet/butt because of the friction from the coarse cement would hurt her & she would still hit the wall. [Prompt: Give at least one piece of evidence from a class activity that supports your ideas] The egg video thingy shows that when momentum is changed very slowly, the stopping force upon said body is also very low. However, if momentum is changed abruptly, the stopping force is very high. Therefore, we know it would be a good idea for the girl to slowly change her momentum, so that her stopping force is small & injuries acquired are also fairly minor.

This is an example of student explanation that is scored as the highest level of sophistication. The observable event (steering slowly toward the grass, drag her feet/bottom, accelerating, hitting the wall, etc.) is clearly distinguished from unobservable mechanism or scientific ideas (changes in momentum, friction) (code = “constructed”). The observable events are linked to unobservable ideas through a coherent chain of cause and effect relationship

(code = “coherent”). The explanation provides a relatively full causal story about various possibilities that the skater girl can choose (code = in-depth “Why”). Finally in terms of use of evidence, not only does the explanation refer to the source of evidence (e.g., the egg video) but the student also thoroughly describes key patterns appearing in the video that support the claim about the suggested skater girl’s choice (code = “evidence-based explanation”).

Example 2—midrange level explanation: A narrated, evidence-referring, and simple “how” explanation with coherence (score = 4 of 7)

I think she should roll on the grass to minimize the force. If she rolls onto the grass the force would spread all over her body so it wouldn’t hurt as much. When you roll on the floor there is friction so it slows you down if you have a lot of ways. [Prompt: Give at least one piece of evidence from a class activity that supports your ideas] roll a ball onto the floor, it will stop at one point.

This is an example of explanation at the midrange level of sophistication, but still above the average scores of the total (average = 2.1). It was coded as “narrated,” not “constructed” because of the connections between observable phenomenon and unobservable science ideas. In the previous example, the changes of speed and motion (observable/conceptual) are clearly discerned from the ideas of momentum and friction (unobservable), and the unobservable concepts were used to describe the mechanism for changes in speed and motion. In contrast, those epistemic distinctions are unclear among the ideas of “roll onto the grass,” “the force would spread all over her body,” and the process of getting injured. This explanation shows a causal coherence beyond describing “what” happens, but not all presented variables are taken into account, therefore, losing the point for “full, gapless in-depth why explanation.” Finally, with respect to the use of evidence, one piece of evidence is cited, but there is no justification for how this pattern supports the proposed claim.

Example 3—low level of sophistication: A narrated, without evidential support and what explanation with no causation (score = 0 of 7)

Because all student responses to the Skater Girl assessment were scored above the average, we pulled out an example of low-level sophistication from the other assessment task. The prompt was “describe how they (a bit of energy) travel through a biological system.”

[Prompt: “Write a story in your journal, pretend you are a bit of energy in this ecosystem. Tell me exactly how you got there, what processes you went through, and how you ended up (a story that will be continued later). Use as many details as you want, turn it into a comic, or make it into a song.”]

Student explanation: I am a sun light molecule, now I am going into this plant through photosynthesis. Now I am going through cellular respiration. I am turned into energy. Then an animal eats me and I am now glucose (sugar) and once again cellular respiration takes place turning me into energy.

In this example, the distinction between the observable phenomenon and unobservable ideas is unclear (code = “narrated”). The explanation is rather explicating a science story than constructing causal relationships (code = “what” explanation, explanation with no causation). There is no support with evidence (code = explanation with no support).

Analytical Approach

The first stage of data analysis focused on understanding descriptively how teachers used scaffolding in designing their assessments. The frequencies of both types and combinations of scaffolding embedded in assessment tasks were calculated. To determine the predictors that explain quality of student explanation, we computed Spearman's correlation coefficients among five types of scaffolding and quality of student explanation. The resulting scatter and box plots indicated a general linear relationship between each of the five scaffolding types and quality of student explanation.

Next, we examined the association between each type of scaffolding and quality of student explanation using hierarchical multiple regression analysis. We intended to use an initial model to examine whether the types of scaffolding would significantly predict the quality of student explanation if other factors were accounted for in advance. Both teacher effects and students' academic background (i.e., students who easily learn, are typical, and are underserved) were controlled as covariates. We created 33 teacher dummy variables and three student dummy variables and entered them in the first and second block of the hierarchical multiple regression analysis, respectively. The five forms of scaffolding were entered in the third block as predictors for quality of student explanations. The following equation describes the model for this hierarchical multiple regression:

$$Y_{\text{quality of explanation}} = \alpha + \beta_0 \times X_t + \beta_1 \times X_s + \beta_2 \times X_1 + \beta_3 \times X_2 + \beta_4 \times X_3 + \beta_5 \times X_4 + \beta_6 \times X_5 + \varepsilon$$

where Y : quality of student explanation, α : an intercept, β_0 : coefficients of 33 teacher dummy variables, β_1 : coefficients of three student group dummy variables, $\beta_2 \sim \beta_6$: coefficients of scaffolding, X_t : 33 teacher dummy variables, X_s : three student dummy variables, $X_1 \sim X_5$: five types of scaffolding, and ε : errors.

This first model helped us examine the general effect of the five types of scaffolding on the quality of student explanation. However, this model did not tell us the effect of each form of scaffolding that was coded as different levels within one type. Four of the five scaffolding types included two different levels of sophistication, such as posterizing (Level 1) versus modeling (Level 2) being levels of the variable "drawing in combination with writing." Furthermore, we were also interested in examining which combinations of scaffolding would best predict the quality of student explanation and the interaction effect between different types of scaffold. Accordingly, we created an additional eight dummy variables for the four scaffoldings that had two different levels and then ran a series of exploratory hierarchical multiple regression analyses. In these analyses, each form (level) of scaffolding used the predictors. A total of 11 models that included two to four types of scaffolding were created. Model 2 shows one example developed to examine the influence of two types of scaffolding combined—drawing and contextualized phenomena.

Model 2: Combinations of two scaffoldings, drawing and contextualized phenomena:

$$Y_{\text{quality of explanation}} = \alpha + \beta_0 \times X_t + \beta_1 \times X_s + \beta_2 \times X_{D1} + \beta_3 \times X_{D2} + \beta_4 \times X_P + \beta_5 \times X_{D1}X_P + \beta_6 \times X_{D2}X_P + \varepsilon$$

where Y : quality of student explanation, α : intercept, β_0 : coefficients of 33 teacher dummy variables, β_1 : coefficients of three student group dummy variables, $\beta_2 \sim \beta_6$: coefficients of scaffolding, X_t : 33 teacher dummy variables, X_s : three student dummy variables, X_{D1} :

Drawing Level 1 (generic drawing or posterizing), X_{D2} : Drawing Level 2 (modeling), X_P : contextualized phenomena, and ε : errors.

Instead of considering all possible combinations, we developed statistical models using the combinations of scaffolding that teachers actually used. This allowed us to examine the effect of actual combinations of scaffolding used in assessments, and those models could be statistically interrogated with empirical data. Interactions among each form of scaffolding were examined in those models, but in the cases of combining more than three types of scaffolding in one model, we put only the interaction term of drawing with other scaffoldings. We made this decision because we were interested in the effect of drawing in combination with other types of scaffolding. Despite its most frequent use, drawing initially appeared as a negative predictor in the result of the overall multiple regression analysis (i.e., Model 1), which puzzled us with regard to the effect of drawing as a form of scaffolding. We also made this decision because of the natural tendency to increase the adjusted R^2 by adding more variables. From these exploratory multiple regression analyses, we first examined the impact of specific categories of scaffolding by comparing the standardized coefficients. Second, we examined which combinations of scaffolding best predict the quality of student explanation by looking at the *patterns* of the adjusted R^2 , keeping in mind the natural increase of R^2 value with the increasing number of variables.

FINDINGS

What Types and Combinations of Scaffolding Do Teachers Use When Designing Explanation-Based Assessment Tasks?

Types of Scaffolding. To address the first research question, we analyzed 76 assessment tasks and found that the first-year science teachers used five types of scaffolding (see Table 3). The most frequently used were (a) encouraging drawing in combination with writing ($n = 42$, 55.3% of the 76 assessment tasks). Following this in frequency of use came (b) contextualizing a phenomenon ($n = 25$, 32.9%), (c) providing a checklist ($n = 21$, 27.6%), (d) providing a rubric ($n = 19$, 25.0%), and (e) providing sentence frames ($n = 10$, 13.1%). As shown in Table 3, only a few assessment tasks provided highly sophisticated forms of scaffolding (i.e., Level 2 scaffolding), such as encouraging drawing a model ($n = 17$, 22.4% of the 76 assessment tasks), providing explanation checklist ($n = 3$, 3.9%), or providing sentence frames that prompt students to make connections ($n = 2$, 2.6%).

Combinations of Scaffolding. The analysis showed that about 20% of assessment tasks did not provide any form of scaffolding ($n = 15$, 19.7%). About 40% of the assessments only provided one type of scaffolding ($n = 29$, 38.2%), and one third provided two or three types of scaffolding ($n = 25$, 32.9%). Less than 8% of assessments provided four types of scaffolding ($n = 6$, 7.9%), and one assessment included all five types of scaffolding ($n = 1$, 1.3%). As shown in Table 4 and indicated by the total frequency of each type of scaffolding in Table 3, providing a space to draw explanations was the most popular type of scaffolding across all the assessments.

Relationship Between Each Type of Scaffolding and Quality of Student Explanation

The results from computing Spearman's correlations showed that all five scaffolding types were statistically significantly correlated with the quality of student explanations. The

TABLE 3
Frequency of the Five Types of Scaffolding Embedded in Assessment Tasks Depending on the Levels of Sophistication

Type of Scaffolding Number of Assessments (%)	Level 0: No Scaffolding Number (%)	Level 1: Moderate Number (%)	Level 2: More Sophisticated Number (%)
Drawing in combination with writing <i>N</i> = 42 (55.3)	No drawing <i>N</i> = 34 (44.7)	Generic drawing/posterizing <i>N</i> = 25 (32.9)	Modeling <i>N</i> = 17 (22.4)
Contextualizing phenomena <i>N</i> = 25 (32.9)	Generic phenomena <i>N</i> = 51 (77.1)	Contextualizing <i>N</i> = 25 (32.9)	NA
Checklist <i>N</i> = 21 (27.6)	No checklist <i>N</i> = 55 (72.4)	Simple words checklist <i>N</i> = 18 (23.7)	Explanation Checklist <i>N</i> = 3 (3.9)
Rubric <i>N</i> = 19 (25.0)	No rubric <i>N</i> = 57 (75.0)	Simple guideline <i>N</i> = 12 (15.8)	Comprehensive table rubric <i>N</i> = 7 (9.2)
Sentence frame <i>N</i> = 10 (13.1)	No sentence frame <i>N</i> = 66 (86.8)	Focusing <i>N</i> = 8 (10.5)	Connecting <i>N</i> = 2 (2.6)

range of the computed Spearman’s rho (ρ) was from .46 to .11. Contextualized phenomena showed the strongest degree of association ($p < .01$, $\rho = .46$) following by the rubric ($p < .01$, $\rho = .41$), checklist ($p < .01$, $\rho = .26$), drawing ($p < .01$, $\rho = .18$), and sentence frame scaffolds ($p < .01$, $\rho = .11$).

The box plots suggested that overall some linear relationships exist between levels of sophistication at each type of scaffolding and the quality of student explanation, with the exception of the rubric (see Figure 4). With respect to the rubric, the quality of student explanation dramatically increased with the Level 1 rubric; in other words, providing simple guidelines as prompts were enough to support students in providing richer scientific explanations. With respect to drawing, engaging students in a generic form of drawing or posterizing (i.e., Level 1) was not significantly related to the quality of student explanation.

What Forms and Combinations of Scaffolding Predict the Quality of Student Explanation?

We wanted to further understand the roles that scaffolding played in the quality of student explanation, controlling for both teacher effect and student academic background. This section presents the results of hierarchical multiple regression analyses using scaffolding as a predictor. The roles of type, combination, and amount of scaffolding are examined both statistically and qualitatively.

The Overall Quality of Student Explanation. The overall average percentage score for quality of the 707 copies of student explanations was 2.1 points of 7, or 30.1% (see Table 5). Among the four dimensions of scientific explanation, the average score in the area of “use of evidence” was particularly low (14.5%). Less than 6% of student explanations (41 of 707) showed strong use of evidence to support students’ explanations.

TABLE 4
Frequency of the Combinations of Scaffolding Used in Assessments

Number of Scaffoldings Used in One Assessment Tasks	Number of Assessments (%)	Combinations of Scaffolding (Number of Assessments)
Five types of scaffolding	1 (1.3)	■ Drawing + contextualizing + rubric + checklist + sentence frame (1)
Four types of scaffolding	6 (7.9)	■ Drawing + contextualizing + rubric + checklist (4) ■ Drawing + contextualizing + sentence frame + checklist (1) ■ Drawing + contextualizing + sentence frame + rubric (1)
Three types of scaffolding	8 (10.5)	■ Drawing + contextualizing + rubric (3) ■ Drawing + rubric + checklist (3) ● Drawing + contextualizing + checklist (2)
Two types of scaffolding	17 (22.4)	● Drawing + contextualizing (6) ● Drawing + rubric (5) ● Drawing + checklist (2) ● Drawing + sentence frame (2) ● Contextualizing + checklist (1) ● Sentence frame + checklist (1)
One scaffold	29 (38.2)	● Drawing (13) ● Contextualizing phenomena (6) ● Checklist (6) ● Rubric (2) ● Sentence frame (2)
No use of scaffolding	15 (19.7)	NA

Using the Five Types of Scaffolding as Predictors for Overall Quality of Student Explanations. We ran hierarchical multiple regression analyses to determine the predictors that explain the quality of student explanations while controlling for teacher and student variances. As described in the preceding section, we developed Model 1 using the five types of scaffolding as the predictors for quality of student explanations, assuming a general linear relationship. The adjusted R^2 changes analysis suggested that about 58% of variance in quality of student explanation were explained by teacher variance, student academic background, and use of scaffolding. Specifically, teacher and student academic background explained about 38% and 9%, respectively. The change of the adjusted R^2 indicated that the use of scaffolding explained about 11% of the variance (see $adj. R^2$ s in Table 6).

Table 6 shows coefficients of the five scaffolding types in Model 1. Three scaffoldings (i.e., contextualized phenomena, providing a checklist, and using rubrics) were positively associated with the quality of student explanation ($p < .05$). Drawing was negatively associated ($p < .05$), and using sentence frames was positive but not statistically significant. The standardized coefficients suggested that contextualizing phenomenon is the strongest predictor of the quality of student explanation ($\beta = .39$), following by rubric ($\beta = .26$) and checklist ($\beta = .19$).

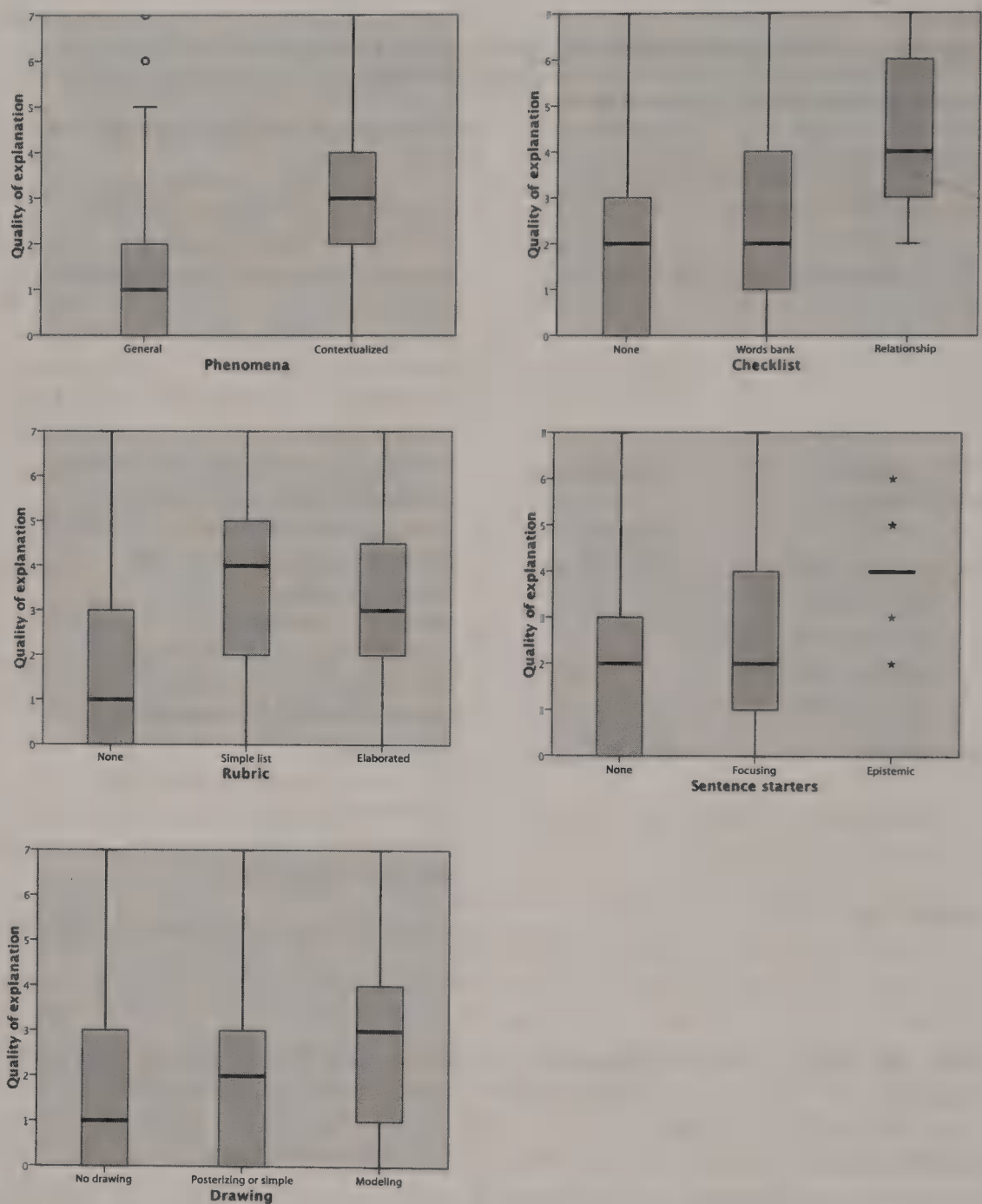


Figure 4. Relationship between each type of scaffolding and quality of student explanation.

TABLE 5
The Average Percentage Scores of Student Explanation

Parameter	Mean Percentage (SD)
Nature of explanation	37.0 (0.48)
Use of evidence	14.5 (0.57)
Depth of explanation	37.5 (0.72)
Coherence of reasoning	35.0 (0.61)
Total	30.1 (1.87)

TABLE 6

Hierarchical Multiple Regression Results for the Impact of Five Types of Scaffolding on Quality of Student Explanation

	<i>B</i>	<i>SE</i>	β	<i>T</i>
First block				
33 Teachers				
	$F(30, 604) = 14.17, p < .001, \text{adj.}R^2 = .38$			
Second block				
Students who are typical	0.66	0.12	.16	5.47***
Students who learn easily	1.34	0.12	.33	11.20***
	$F_{\text{change}}(2, 602) = 52.93, p < .001, R^2_{\text{change}} = .09$			
Third block				
Contextualizing phenomena	1.55	0.17	.39	9.31***
Rubric	0.75	0.12	.26	6.35***
Checklist	0.65	0.13	.19	5.17***
Drawing	-0.49	0.13	-.20	-3.75***
Sentence frame	0.13	0.17	.03	.77
	$F_{\text{change}}(5, 597) = 30.78, p < .001, R^2_{\text{change}} = .10$			
The overall model: $F(37, 597) = 24.52, p < .001, \text{adj.}R^2 = .58$				

Notes: ^aA total of 36 dummy variables were created with respect to 33 teachers and three groups of students (students who learn easily, are typical, and underserved). Underserved students are reference group (coded as 0). The coefficients for 33 teachers are not reported in this table for its brevity.

*** $p < .001$.

Using Each Form of Scaffolding as a Predictor for Quality of Student Explanations.

The first purpose of these analyses was to examine the effect of different levels of *sophistication* within a type of scaffolding (e.g., posterizing vs. modeling, simple checklist vs. explanation checklist). The second purpose was to compare the effect of different *combinations* of scaffolding (i.e., combination of two or more forms of scaffolding). Finally, we intended to examine the interaction effect among scaffolding types, particularly with drawing.

Examination of the standardized coefficients of each form of scaffolding across 11 regression models suggested that using contextualized phenomena is the strongest single predictor of the quality of student explanation. This result is consistent with the result from the previous regression analysis (Model 1). Both rubric and checklist were also significant predictors, as suggested in Model 1. A Level 2 checklist (i.e., explanation checklist) is a stronger predictor than a Level 1 checklist (i.e., simple word list). Interestingly, however, a Level 1 rubric (i.e., simple rubric) is a stronger predictor than a Level 2 rubric (i.e., elaborated rubric).

In terms of the effect of combinations of scaffolding, the most powerful combination was the one that had three or more upper level scaffolding types that *included contextualized phenomena*. Most statistical models from the combinations of three or four scaffoldings show consistently higher values of the adjusted R^2 (.56–.57). In contrast, using two scaffoldings showed some spectrum of the adjusted R^2 from .50 to .55, indicating that the selection of forms of scaffolding is critical. The following section features a case of a teacher using a task with three upper level scaffolds (*contextualizing a phenomena, drawing as modeling, and a comprehensive rubric*) and one low level scaffold (a “focusing” sentence frame). The case illustrates the nature of opportunities created with use of scaffolding for student

Name _____ Date 10/25/11 Period 5

FINAL EXPLANATION **WHY DON'T COUNTRIES NEAR THE EQUATOR, LIKE SAMOA, SEEM TO HAVE SEASONS LIKE WE DO HERE, IN SEATTLE?**

1. Draw four Earths that show what causes Seattle to experience:
 a. Fall b. Winter c. Spring d. Summer +4

2. Label each Earth with the season.

3. Label Seattle and Samoa (or another country near the Equator that you've chosen, like Cambodia or Kenya) on your drawing in each Earth.

(6 pts)

4. Write a claim here that answers the guiding question. (2 pt) +1

Countries near the Equator, like Samoa, don't seem to have seasons like we do here, in Seattle because... the Northern hemisphere is tilted towards the Sun (Summer in Seattle, Samoa never really experiences seasonal changes because Samoa receives 48% but why does Samoa never really experience seasons)

5. Provide evidence from activities in class (use your summary table) to support your answer and explain how your evidence supports your answer using your drawing. Change

Look at the rubric to help you decide what information to include. (6 pts) +4

The solar radiation hits the Equator at a 90 degree angle. The Earth's seasons are caused by its rotation and tilt of the Earth around the Sun. Because the Earth's tilt causes the Northern and Southern hemispheres to receive more intense direct solar radiation during part of the year in summer, we only see daylight and temperature patterns associated the seasons. The Earth is tilted. In our Solar Cell demo, the solar cell spins slower when the light has to go through more sheet protectors, so that means that tilt slows down filters solar radiation, diffusing it. The Earth's tilt causes more solar radiation at the equator. In our solar cell demo, the solar cell spins faster the closer it is to the lamp.

Figure 5. Seasonal change assessment: Using a combination of three high- and one low-level scaffolding in a task.

responses, and how the teacher used this assessment to further support and enhance student learning.

Seasonal Change Assessment: Using a Combination of Three High- and One Low-Level Scaffolding in a Task. This focal assessment was about seasonal changes in a seventh-grade earth science unit. Students were prompted to explain: “Why don’t countries near the equator, like Samoa, seem to have seasons like we do here, in Seattle?” As shown in Figure 5, students drew the positions of the earth around the sun in the four seasons and then labeled Seattle and Samoa on their drawing. The teacher gave students options to choose another country near the equator like Cambodia or Kenya. It should be noted that this school has a large population of students of immigrant families from these regions of the world. Next, students were prompted to write a claim in a few sentences guided by the sentence frame: “Countries near the equator, like Samoa, don’t seem to have seasons like we do here, in Seattle because . . .” It follows the prompt of “Providing evidence from activities in class (use Summary Table²) to support your answer and explain how your evidence supports your answer using your drawing. Look at the rubric to help you decide what information to include.” By combining modeling with contextualization in this assessment, students were invited to engage in a high level of intellectual work that involved (a) locating geographic positions for two different countries on the earth, (b) identifying the relative positions of the sun and earth during the orbit of the earth at different seasons, and (c) simulating the seasonal changes at two different locations in terms of exposure to the sun’s light. The *focusing sentence frame* seems to help students to get right into the heart of the work, that is, writing a claim about the focal phenomena. In this assessment task, the combination of this high-level scaffolding enabled students not only to draw on their everyday reasoning resources, such as the seasonal differences noted by themselves

² A Summary Table is a form of public representation that lists activities (labs, readings, demonstrations) and key ideas addressed with each of the activities, in the form of a table.

and their relatives and travel experiences to visit their relatives, but also to express their ideas in a modality other than writing.

Most students demonstrated significant progresses in their explanation about seasonal changes. For example, one student, Nick, a student as a typical category, initially had a “distance theory” about seasonal changes—“the summer is hot because the earth is closer to the sun, and winter is cold because the earth is far.” This was Nick’s initial idea about seasonal changes that were elicited on the first day of this unit. Nick produced the following explanation 2 days before the final day of this unit as response to the assessment task:

[Sentence frame] Countries near the Equator, like Samoa, don’t seem to have seasons like we do here, in Seattle because . . . [Nick’s response] the Northern hemisphere is tilted towards the sun (Summer in Seattle). The solar radiation hits the equator at a 90 angel [sic]. The Earth’s seasons are caused by the rotation and tilt of the earth around the sun. Because the Earth’s tilt causes the Northern and Southern Hemispheres to receive more intense/direct solar radiation during part of the year in summer. We only see daylight and temperature patterns associated the seasons if the Earth is tilted. In our Solar Cell demo, the solar cell spins slower when the light has to go through more sheet protectors. So that means that the atmosphere filtered solar radiation, diffusing it. (score: 6 of 7, a constructed, evidence-based, in-depth why explanation with partial coherence & reasoning)

The solar cell investigations being referred to were designed to show that not only is the sun’s light more concentrated in northern latitudes during the months of June through August (amount of radiant energy per unit area), but that when the sun’s light is at a less oblique angle to the earth’s surface, it does not have to pass through as many miles of atmosphere before reaching the surface. As shown in Figure 5, the teacher provided feedback to this question in the form of question to press students to further elaborate his idea: “Yes, but why does Samoa never really experience seasonal changes?” Nick revised his model using a different color pen, stating, “Samoa never really experiences seasonal changes because Samoa receives more solar radiation and stays like that all the year long.” He also added his evidence about it: “The Earth’s tilt causes more solar radiation at the equator. In our solar cell demo, the solar cell spins faster the closer it is to the lamp.”

The overall average score of student explanations across 12 samples of student work from this assessment was 4.4 of 7 (63.1%). Of note, the four students who were identified as the category of “learns with difficulty” did as well as the four students in the typical group (the average score of students in easily learn group: 6.3, typical 3.5, and learn with difficulty: 3.5).

DISCUSSION

We make two claims from this study. First, with effective use of scaffolding, teachers create better opportunities for students to demonstrate disciplinary proficiency. Some types of scaffolding, such as using explanation checklists and contextualizing phenomena, clearly define the level of rigor expected of students and appear to make the task more accessible for learners. Second, the *quality and combination* of scaffolding types matters more than the number of scaffolds embedded in the assessments. Quality combinations include a contextualized phenomenon in addition to one or more of the other types of upper-level scaffolds. To unpack these claims, we first focus on how each type of scaffolding supports students’ construction of evidence-based explanations. Then, we discuss how and why combinations of high-quality scaffolding are particularly effective. We finish this section

with a discussion of the relationships among scaffolding, teachers' instruction, and student learning.

Role of Each Type of Scaffolding in Supporting Evidence-Based Explanation

Contextualized Phenomena: Making the Task Intellectually Challenging But Accessible. We hypothesize that contextualization helps students engage in deeper forms of reasoning and demonstrate in-depth explanations in four ways. First, contextualization problematizes a generic set of conditions. For example, in the assessment about the seasons, contextualizing the generic phenomena of seasonal changes in two different geographic regions (i.e., Samoa and Seattle) generates multiple variables that must be taken into account, such as relative distances of different parts of the earth from the sun (negligible), changes in the angles of the sunlight on the earth, penetration of light through layers of the atmosphere, etc. This contextualization prompts students to recognize the general model of seasonal change and then reason how this model plays out under different conditions. Second, a contextualized phenomenon supports students in moving beyond reproductions of textbook explanations about general phenomena. The fact that there is no authoritative "answer" from a textbook helps a teacher and students reposition themselves as coinvestigators in the process of actually making sense of the phenomena. Third, situating the phenomena in the everyday experiences of students and their families helps them draw in additional intellectual resources from observations and relevant accounts of others (Nordine, Krajcik, & Fortus, 2010; Shwartz, Weizman, Fortus, Krajcik, & Reiser, 2008). For example, in the skater girl assessment (see Figure 3), students used scientific ideas, such as momentum, friction, and acceleration, but also drew upon their everyday reasoning resources, such as "she shouldn't drag her feet/bottom because of the friction from the coarse cement would hurt her" or "When you roll on the floor there is friction so it slows you down if you have a lot of ways." In this way, students drew on prior experiences and observations about slowing down as their bodies interacted with different surfaces and made sense of science ideas simultaneously. Finally, contextualizing a phenomenon in a particular local community helps students relate to the problem, which allows them to become engaged in the work actively and emotionally. Walqui (2006) describes the importance of "bridging" as establishing a personal link between the students and the subject matter by showing the relevance of new materials to the students' life "here and now" (p. 172). The problems that produced richer student explanations—richer meaning longer, more detail, and support for a claim—tended to describe an event/phenomenon that was relevant to the students' everyday life, such as the skater girl's story. In short, contextualized phenomena makes the task more cognitively challenging by problematizing a generic set of conditions and by inviting students to engage in complex reasoning, but at the same time its situatedness in a set of recognizable conditions makes the task more accessible to students.

Providing ■ Rubric: Priming Disciplinary Ways of Thinking and Talking. Rubrics are frequently used to engage students in monitoring their current level of thinking and determining next steps toward learning goals (Shepard, 2005; Walqui, 2006). In this study, rubric scaffolds alone were significantly associated with the quality of student explanations. Many rubrics explicitly encouraged epistemic features of disciplinary thinking and talking. Even prompts at a Level 1 revealed the tacit structure of evidence-based explanations to students. For example, in the skater girl assessment, each prompt guided students to describe what happened (i.e., a skater girl who is about to hit the wall), use scientific concepts as part of

the explanation (i.e., force, friction, momentum) and then to describe the reasoning behind a claim (i.e., explain why this is better than another choice she has). Shepard (2000) reminds us that the importance of transparency is to make the assessment an activity for learning. A rubric that makes the expectations for achievement transparent seemed to support students in successfully constructing evidence-based explanations.

Providing Checklists: Inviting Complex Reasoning With Multiple Relationships.

Checklists, in particular an explanation checklist that consisted of statements about important relationships relevant to the focal phenomena, significantly explained the quality of student explanations. We hypothesize that even a simple word checklist reduces the cognitive load for locating terminology, and refocuses students' intellectual resources toward synthesizing information, examining relationships, and evaluating evidence. In contrast, more elaborated forms of checklists, such as explanation checklists, highlight particular relationships or dimensions of process that students need to consider. For example, in the cell membrane assessment (see Figure 2), the explanation checklist helped students attend to key relationships that define a system, such as paramecium's in-take of water in relation to the amount of salt. Also, it invites students to reason about how the relationship fits into a larger activity system (e.g., paramecium's survival). Without explanation checklists, students were able to produce explanations, but with this form of checklist the task challenged students to consider specific dimensions of the phenomena that they may not have otherwise attended to in written explanations.

Drawing in Conjunction With Writing: Prompting to Attend to Relationships and Underlying Mechanisms. The combination of drawing with other forms of scaffolding showed substantially different and positive impacts of two different forms of drawing—namely, generic drawing or posterizing (i.e., Level 1) in contrast with modeling (Level 2)—as predictors. When students engaged in posterizing, they illustrated known facts or information from authoritative sources, typically from textbooks. For example, in the rock cycle assessment task in an earth science unit (see Figure 1), most students produced almost the same drawings that simply illustrated the scientific model of the rock cycle. This kind of drawing may direct students to reproduce the canonical scientific models rather than actively engaging in sensemaking processes. In contrast, when students were prompted to construct models that were products of their own sense making through drawing, they produced diverse types of inscriptions that revealed a wider variety of developing ideas, partial understandings, and different ways of reasoning. For example, in the assessment about buoyancy (see Figure 1), students were prompted to first draw a person floating in salt water, then in fresh water, and to describe all the forces acting on the person, and then to write their explanation of how the density of a fluid affects the buoyant forces. Engaging in this assessment task, students used their drawings to explain the focal phenomenon (i.e., why people float higher in salt water than in fresh water) by highlighting the underlying mechanism (i.e., how the density of a fluid affects buoyant forces in relation to the force of gravity).

We theorize that “drawing as modeling” provides support for student reasoning in the process of making sense of relationships among events, structures, properties, and concepts. Modeling prompts students to identify unobservable events/processes and then connect them causally to patterns of observation (Windschitl, Thompson, & Braaten, 2008). The prompts to illustrate changes either over time or between conditions guide students to express the relationship between unobservable events and changes in the state of the

systems across conditions. All of these reasoning tasks are challenging for learners new to reasoning with scientific practices.

Previous studies have suggested that scaffolding different modes of students' expression of ideas, such as drawing, open up opportunities for students who may have difficulty in using scientific language. This is one of the major barriers in science learning (Moje et al., 2001; Rosebery et al., 2010). Our analyses suggest that drawing allows students to show more of what they know, but it needs to be prompted in particular ways to support students' representations of evidence-based explanations.

Sentence Frames: Guiding Disciplinarily Valid Ways of Thinking and Talking. We conjecture that well-designed sentence frames support disciplinary and epistemic reasoning. They also help students express their thinking semantically. "Focusing" sentence frames (i.e., Level 1) usually took the form of "things happened because ____" and provided a semantic structure to construct causal explanations about the focal event. Previous studies have shown that students often have difficulty addressing the main question or issue when prompted for explanations (e.g., Ruiz-Primo et al., 2010). By providing a "focusing" sentence frame, students are guided into the problem space. Furthermore, more sophisticated sentence frames, specifically "connecting" sentence frames, provided additional support for using epistemic structures for writing scientific explanations. As demonstrated in the cell membrane assessment (Figure 2), sentence frames guided students to construct their explanation while examining the relationship between evidence and reasoning. Providing clear examples of appropriate language use for the performance of specific academic functions, such as the use of evidence, is a critical form of scaffolding (Walqui, 2006). It seems that well-designed sentence frames have potential to support the epistemic and semantic challenges of constructing evidence-based explanations. In our study, however, teachers used sentence frames infrequently ($n = 10$, 13.1%), and upper level sentence frame were rarely used. More evidence is needed to understand the roles of different forms of sentence frames in supporting students' productive disciplinary participation.

Quality and Combinations of Scaffolding Matter

The results of hierarchical multiple regression analyses show no linear relationship between the number of scaffolding types used in assessments and the quality of student explanations. In other words, simply adding more forms of scaffolding did not increase the quality of student explanations. However, close examination of both quantitative and qualitative analyses suggests that when several forms of *higher level scaffolding types* were used in *strategic combinations*, it creates opportunities for more students across different academic achievement levels to show a spectrum of ideas and reasoning through written explanation.

Constructing evidence-based explanations is a highly complex task that poses multidimensional challenges in understanding the task itself, planning a response, and producing representations of one's thinking. Different learners encounter different kinds of challenges in the process of constructing evidence-based explanations. Students, for example, vary in their reading and writing proficiencies as well as their self-perceptions as science learners. As discussed in the preceding section, each form of scaffolding serves its own unique function in supporting the work of constructing evidence-based explanations while addressing different challenges for such performances. Therefore, a combination of several types of high-level scaffolding can create opportunities for more students to succeed in engaging in constructing evidence-based explanations.

Scaffolding, Teachers' Classroom Instruction, and Supporting Deeper Learning

The assessment tasks in this study were designed to be used as a part of teaching. We noticed that the most sophisticated forms of assessment were developed by the teachers who were known to enact similarly sophisticated classroom practices. We conjecture that the teachers who created quality opportunities for students to demonstrate their ideas, thinking, and ways of reasoning with the effective use of scaffolding come to be in a better position to modify their teaching practices because the scaffold-rich assessments produce highly descriptive information about students' strengths and difficulties. This study focused on one critical part of assessment design—providing opportunities for students to show what they are capable of.

Our findings implicate the intertwined relationship between the performative aspects of teaching practices and the use of scaffolding. Students' written explanation produced from their engagement in assessment tasks is an outcome of activity that is always situated in larger instructional contexts. It is reasonable to ask to what degree the increased quality of student explanations is the effect of the embedded scaffolding within assessments. From hierarchical multiple regression analyses, the teacher—as a variable—explained the greatest amount of variance in student performance (about 38%), followed by use of scaffolding (about 11%) and students' academic backgrounds (about 9%).

The findings strongly indicate a clear relationship between students' explanatory performance and the use of scaffolding. In our data set of 707 copies of student work, the overall score for quality of student explanation was 2.1 of 7 (30.1%). This result indicates that most students are currently failing to meet the expectations grounded in ambitious visions of twenty-first century learning. In a recent study that examined the quality of students' written explanations from science inquiry-based middle school classrooms in five states, Ruiz-Primo et al. (2010) also found that a low percentage of students (18%) provided explanations meeting the criteria for evidence-based explanation. A close examination of teachers' use of scaffolding and student performance in the present study suggests that the lower level of student performance has to do with teachers' less frequent use of scaffolding. In this study, about one quarter of assessment tasks did not provide any form of scaffolding (15 of 76 assessment tasks, 19.7%). In those assessments, the average quality of student explanation was about half ($n = 121$, average score = 1.12 of 7; $SD: 1.10$) compared to student responses from those with *any form* of scaffolding ($n = 586$, average score = 2.32 of 7; $SD = 1.94$). Even when the assessment tasks provided some scaffolding, those were generally one or two types of scaffolding with *lower levels* of sophistication.

Most assessment tasks did not scaffold students' substantial use of evidence for scientific explanations. It should be noted that new forms of learning that we are envisioning, such as constructing evidence-based explanations, involve students' engagement in particular ways of thinking and talking that are unfamiliar to most students. Ruiz-Primo and her colleagues (2010) noted that

[w]hat students are missing and lacking is learning experiences and guidance in the fundamental activities of constructing explanations. Unless or until such experiences and guidance are adequately offered, it is not surprising to find that constructing explanations is challenging for students. (p. 605)

It would be unrealistic for our students to meet the increased expectations without providing appropriate and sufficient support. Put another way: Designing for rigor in curriculum in student learning experiences or in assessments may not necessarily lead to higher student

performances. A narrow focus on rigor may in fact obscure what students are capable of, if given strategic (and temporary) forms of scaffolding.

This leads to the issue of fading. The question of fading supports while simultaneously maintaining necessary support of all students' efforts at science reasoning and practice remains understudied. The big concern is "when" and "under which conditions" will students be able to construct evidence-based explanations without scaffolds? We hypothesize that the absence of scaffolds of any kind will be most challenging when students are presented with a novel phenomenon that is neither directly relevant to students' everyday lives nor locally contextualized. We also believe that, across units of instruction, some scaffolds may gradually be removed as classroom communities become more practiced in talking and writing about models and explanations. In terms of further research, it will be fruitful to take a situative perspective (Greeno, 2006) to understand the varied ways students learn to appropriate the habits of thinking behind each form of scaffolding and what kinds of learning environments are most influential for students to take up particular kinds of writing and discourse. We recognize that there cannot be a one-size-fits-all approach to fading. But we may be better able to understand the process of fading by showing how scaffolds for varied learners are released over time and observing these effects in the contexts of the development of social and scientific communities in K-12 classrooms.

CONCLUSION AND IMPLICATIONS

We conclude that providing effective scaffolding is necessary, not optional, when trying to support students in meeting twenty-first century standards (NRC, 2012a; Resnick, 2010). For teachers and science teacher educators who are interested in designing assessment tasks for supporting and enhancing student learning, we recommend using a combination of two or more high-quality types of scaffolding including the use of contextualized phenomena. This use of contextualized phenomena has implications, of course, for instruction as well as assessment. The use of "anchoring" phenomena for units of instruction has been strongly suggested in the Framework document for the Next Generation Science Standards (NRC, 2013).

With the new reform emphases on science as practice, we see an emerging challenge for teachers—that of scaffolding both instruction around these practices and the assessment of students' abilities to take up the conceptual, social, and epistemic dimensions of disciplinary work. We believe that assessment and instruction must be interwoven more than they currently are and not treated as separate events. If classroom teachers can "see" more of what their students are capable of through well-designed assessments, there will be new opportunities to study how they use these rich forms of data to make instructional decisions that attend to the needs of different groups of learners. With scaffolding, the quality of information to base decisions on is much greater than without. We see this as a lever to promote both rigor and equity in classroom teaching.

We wish to thank the teachers who provided their student work and were willing to share their stories with us. We also thank to the research group at the University of Washington for their help in developing this paper. This article benefitted from the critical feedback of three anonymous reviewers and the editor of the journal.

REFERENCES

- Braaten, M., & Windschitl, M. (2011). Working toward a stronger conceptualization of scientific explanation for science education. *Science Education*, 95(4), 639–669.

- Bransford, J. D., Brown, A. L., & Cocking, R. R. (1999). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academies Press.
- Brown, A. L., & Campione, J. C. (1994). Guided discovery in a community of learners. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice*. Cambridge, MA: MIT Press/Bradford Books.
- Bruner, J. (1985). Narrative and paradigmatic modes of thought. In E. Eisner (Ed.), *Learning and teaching: The ways of knowing*. Chicago: National Society for the Study of Education.
- Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks, CA: Sage.
- Ford, M. J., & Wargo, B. M. (2012). Dialogic framing of scientific content for conceptual and epistemic understanding. *Science Education*, 96(3), 369–391.
- Furtak, E. M., & Ruiz-Primo, M. A. (2008). Making students' thinking explicit in writing and discussion: An analysis of formative assessment prompts. *Science Education*, 92(5), 799–824.
- Greeno, J. G. (2006). Learning in activity. In K. R. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 79–96). New York: Cambridge University Press.
- Herman, J. L. (1992). What research tells us about good assessment. *Educational Leadership*, 49(8), 74–78.
- Kenyon, L., & Reiser, B. J. (2006). A functional approach to nature of science: Using epistemological understandings to construct and evaluate explanation. Paper presented at the AERA annual meeting, San Francisco, CA.
- Knorr-Cetina, K. (1999). *Epistemic cultures: How sciences make knowledge*. Cambridge, MA: Harvard University Press.
- Latour, B., & Woolgar, S. (1979). *Laboratory life: The social construction of scientific facts*. Los Angeles: Sage.
- McNeil, K. L., & Krajcik, J. (2006). Supporting students' constructions of scientific explanation through generic versus context-specific written scaffolds. Paper presented at the AERA Annual Meeting, San Francisco, CA.
- McNeil, K. L., & Krajcik, J. (2008). Scientific explanations: Characterizing and evaluating the effects of teachers' instructional practices on student learning. *Journal of Research in Science Teaching*, 45(1), 53–78.
- Moje, E. B., Ciechanowski, K. M., Kramer, K., Ellis, L., Carrillo, R., & Collazo, T. (2004). Working toward third space in content area literacy: An examination of everyday funds of knowledge and discourse. *Reading Research Quarterly*, 39, 38–72.
- Moje, E. B., Collazo, T., Carrillo, R., & Marx, R. W. (2001). "Mastro, what is 'quality'?"?: Language, literacy, and discourse in project-based science. *Journal of Research in Science Teaching*, 38(4), 469–498.
- Nasir, N. S., Rosebery, A. S., Warren, B., & Lee, C. (2006). Learning as a cultural process: Achieving equity through diversity. In K. R. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 489–504). New York: Cambridge University Press.
- National Research Council. (2000). *Inquiry and the National Science Education Standards*. Washington, DC: National Academy Press.
- National Research Council. (2007). *Taking science to school: Learning and teaching science in grade K-8*. Washington, DC: National Academy Press.
- National Research Council. (2012a). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Washington, DC: The National Academies Press.
- National Research Council. (2012b). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academies Press.
- National Research Council. (2013). *Next generation science standards: For states, by states*. Washington, DC: The National Academies Press.
- Nersessian, N. (2005). Interpreting scientific and engineering practices: Integrating the cognitive, social, and cultural dimensions. In M. Gorman, R. D. Tweney, D. Gooding, & A. Kincannon (Eds.), *Scientific and technological thinking* (pp. 17–56). Hillsdale, NJ: Erlbaum.
- Nordine, J., Krajcik, J., & Fortus, D. (2010). Transforming energy instruction in middle school to support integrated understanding and future learning. *Science Education*, 95(4), 670–699.
- Ohlsson, S. (2002). Generating and understanding qualitative explanations. In J. Otero, A. Leon, & A. C. Graesser (Eds.), *The psychology of science text comprehension*. Mahwah, NJ: Erlbaum.
- Osborne, J., & Patterson, A. (2011). Scientific argument and explanation: A necessary distinction? *Science Education*, 95(4), 627–638.
- Pea, R. D. (2004). The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity. *Journal of the Learning Sciences*, 13(3), 423–451.
- Pickering, A. (1995). *The mangle of practice: Time, agency, and science*. Chicago: The University of Chicago Press.
- Resnick, L. B. (2010). Nested learning systems for the thinking curriculum. *Educational Researcher*, 39(3), 183–197.

- Rosebery, A. S., Ogonowski, M., DiSchino, M., & Warren, B. (2010). "The coat traps all your body heat": Heterogeneity as fundamental to learning. *Journal of the Learning Sciences*, 19(3), 322–357.
- Ruiz-Primo, M. A., & Furtak, E. M. (2007). Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. *Journal of Research in Science Teaching*, 44(1), 57–84.
- Ruiz-Primo, M. A., Li, M., Tsai, S.-P., & Schneider, J. (2010). Testing one premise of scientific inquiry in science classrooms: Examining students' scientific explanations and student learning. *Journal of Research in Science Teaching*, 47(5), 583–608.
- Sandoval, W. A. (2003). Conceptual and epistemic aspects of students' scientific explanations. *Journal of the Learning Sciences*, 12(1), 5–51.
- Sandoval, W. A., & Millwood, K. A. (2005). The quality of students' use of evidence in written scientific explanations. *Cognition and Instruction*, 23(1), 23–55.
- Sandoval, W. A., & Reiser, B. J. (2004). Explanation-driven inquiry: Integrating conceptual and epistemic scaffolds for scientific inquiry. *Science Education*, 88(3), 345–372.
- Sawyer, K. R. (2006). The new science of learning. In K. R. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 1–16). New York: Cambridge University Press.
- Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, 29(7), 4–14.
- Shepard, L. A. (2005). Linking formative assessment to scaffolding. *Educational Leadership*, 63(3), 67–70.
- Shwartz, Y., Weizman, A., Fortus, D., Krajcik, J., & Reiser, B. (2008). The IQWST experience: Using Coherence as a Design Principle for a Middle School Science Curriculum. *The Elementary School Journal*, 109(2), 199–219.
- Smith, C. L., Maclin, D., Houghton, C., & Hennessey, M. G. (2000). Sixth-grade students' epistemologies of science: The impact of school science experiences on epistemological development. *Cognition and Instruction*, 18(3), 349–422.
- Songer, N. B., & Gotwals, A. W. (2012). Guiding explanation construction by children at the entry points of learning progressions. *Journal of Research in Science Teaching*, 49(2), 141–165.
- Supovitz, J. (2012). Getting at student understanding—The key to teachers' use of test data. *Teachers College Record*, 114, 1–29.
- Thompson, J., Braaten, M., Windschitl, M., Sjöberg, B., Jones, M., & Martinez, K. (2009). Collaborative inquiry into students' evidence-based explanations: How groups of science teachers can improve teaching and learning. *The Science Teacher*, November, 48–52.
- Toulmin, S. (1958). *The uses of arguments*. Cambridge, England: Cambridge University Press.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
- Walqui, A. (2006). Scaffolding instruction for English language learners: A conceptual framework. *The International Journal of Bilingual Education and Bilingualism*, 9(2), 159–180.
- Warren, B., Ballenger, C., Ogonowski, M., Rosebery, A. S., & Hudicourt-Barnes, J. (2001). Rethinking diversity in learning science: The logic of everyday sense-making. *Journal of Research in Science Teaching*, 38(5), 529–552.
- Windschitl, M., Thompson, J., & Braaten, M. (2008). Beyond the scientific method: Model-based inquiry as a new paradigm of preference for school science investigations. *Science Education*, 92(5), 941–967.
- Windschitl, M., Thompson, J., & Braaten, M. (2011). Ambitious pedagogy by novice teachers: Who benefits from tool-supported collaborative inquiry into practice and why? *Teachers College Record*, 113(7), 1311–1360.
- Windschitl, M., Thompson, J., Braaten, M., & Stroupe, D. (2012). Proposing a core set of instructional practices and tools for teachers of science. *Science Education*, 96(5), 878–903.

Long-Term Self-Regulation of Biology Learning Using Standard Junior High School Science Curriculum

BILLIE EILAM, SHOSHI REITER

Faculty of Education, University of Haifa, Mt. Carmel, Haifa 31905, Israel

Received 27 February 2013; accepted 2 April 2014

DOI 10.1002/sce.21124

Published online 18 June 2014 in Wiley Online Library (wileyonlinelibrary.com).

ABSTRACT: In today's world of information explosion, independent lifelong self-regulated learning (SRL) is becoming a necessity. However, opportunities in schools to experience such learning modes are relatively rare. This long-term explorative field study examined students' SRL of science. Changes in students' self-reported and enacted SRL in two ninth-grade science classrooms were measured over the course of a full academic year. The self-regulating (SR) class ($n = 25$) studied genetics while self-regulating their learning. The teacher-controlled (TC) class ($n = 27$) studied the same content in a teacher-regulated mode. Data were collected at several time points along the year by SRL questionnaires (Learning and Study Strategies Inventory, LASSI), science knowledge tests, and (in the SR group only) protocols for measuring enacted SRL. Findings showed that enacted SRL correlated with self-reported SRL, validating measurements. The SR students outperformed their TC peers in science learning. Significant changes over time in both self-reported and enacted SRL emerged only for the SR students, suggesting that long-term SRL practices may promote awareness of learning processes and ability to apply SRL. Implications for school learning are discussed. © 2014 Wiley Periodicals, Inc. *Sci Ed* 98:705–737, 2014

INTRODUCTION

The goals of the national high school science curriculum in Israel and most Western countries have been significantly affected by changes occurring in today's science- and technology-oriented society. Such changes include the information explosion and changing

Correspondence to: Billie Eilam; e-mail: beilam@edu.haifa.ac.il

job markets, which require today's citizens to become independent and educated lifelong learners (Schraw, Crippen, & Hately, 2006). These societal changes are manifested in the Israeli science curriculum's emphasis on two main goals: the construction of science knowledge and the acquisition of relevant skills, particularly the skills needed for initiating a shift in students' behavior from relying on teacher control to becoming autonomous learners (Roth & Lee, 2004). Learner autonomy in our case refers to student ability to make decisions with regard to all aspects of their learning processes, including pace, activities, setting, and so forth. Such learning processes constitute the basis for lifelong learning through self-regulation. Nevertheless, self-regulated learning (SRL) is not yet a common practice in schools (Cleary & Labuhn, 2013; Van Velzen, 2013).

The ability for SRL is not a trait, and although individual differences in SRL levels and styles have been reported; research studies have shown that all students, regardless of their academic abilities, can improve their SRL and, in turn, their academic achievement (Kiewra, 2002; Perels, Dignath, & Schmitz, 2009). Its acquisition benefits both individuals and society as a whole because SRL promotes thinking skills and learning behaviors and, therefore, may decrease student academic diversity (Dignath, Buettner, & Langfeldt, 2008). Moreover, SRL was found to correlate negatively with student dropout, especially at the beginning of junior high school (Van der Veen & Peetsma, 2009). Regarding science in particular, effective learning requires the nurturing of SRL as essential for meeting the increasing demands of present-day science education goals (Lee, Lim, & Grabowski, 2009; Schraw et al., 2006).

Such positive evidence of the benefits of SRL calls for organizing plentiful opportunities within the framework of school curricula to enable student regulation of their own learning processes. These opportunities should involve two main conditions: (1) They should be long-term experiences and provide ample time for student reflection, monitoring, trials, failures, consideration, and recovery without suffering serious repercussions (Eilam, Zeidner, & Aharon, 2009); (2) they should involve the regulation of complex tasks so that students may notice the relations between goals, strategies, products, and the environment, thereby constructing and refining their mental models of both the content studied and these SRL processes (Eilam & Aharon, 2003). Essential opportunities like these require the design of a unique learning environment to empower students' autonomous learning.

The present study implemented such a unique environment while investigating ninth graders' yearlong online self-reported and enacted SRL of the national genetics curriculum and their outcomes in terms of constructed knowledge. It focused on changes in students' regulatory behaviors over time and in their academic performance in this autonomous learning environment as compared with students learning in a teacher-controlled (TC) environment. For this purpose, we used an instrument for promoting students' SRL and recording their regulating performance for analysis.

STUDY FRAMEWORK

In this section, we shall discuss the theoretical concepts and issues applied in our study, beginning with the difficulties inherent to learning biology.

Difficulties in Science and Biology Learning

Biology learning may evoke many difficulties for students as it requires high cognitive flexibility as well as diverse information processing skills. For example, to construct a deep understanding of biological systems, students need to shift flexibly between the macro- and microlevels of organism structure and function, be aware of the interactions between

these levels, and link domains to their subdomains (e.g., linking cytology to genetics) (Bronsan, 1990; Lee et al., 2009; Lewis & Kattman, 2004). Other difficulties emerge from students' entrenched misconceptions and their deficiency in high-order thinking skills, which may hinder biology learning (Eilam, 2012a; Lewis & Kattman, 2004). Research has also pinpointed students' difficulties in learning with models and visual representations (e.g., equations, three-dimensional [3D] cells, or DNA structures represented on a two-dimensional [2D] space), which are common and essential in biology instruction (Eilam, 2012b; 2012c). Biological concepts may be concrete or abstract and may conceptualize phenomena of any scale (e.g., the biosphere, cell, DNA). Owing to the systemic nature of biology content, the construction of deep understanding of a single concept may require knowledge of other, related concepts. In addition, the need to integrate information from sources of different modalities to construct understanding (e.g., physical-laboratory, visual-textual) was shown to challenge learners due to the need to map across these different resources (Eilam, 2012b; Lemke, 2002).

Science textbook content, for example, is presented to students linearly and sequenced to fit scientific logic. In most cases, textbooks remain a teacher's major pedagogical device and students' primary information resource (Apple, 2008). However, students do not necessarily think linearly, and their own bodies of knowledge may be organized differently from the presented organization (Stinner, 1992). To overcome some of these difficulties, researchers advocate the active involvement of students in learning (Lin & Eylon, 2006) and in the regulation of their own learning processes (Eilam, 2012a; Lee et al., 2009). Therefore, in the present intervention students engaged deeply in the process of constructing their biology knowledge and reorganizing their learning environment to fit to their own personal preferences in light of situational difficulties encountered (e.g., preferring collaboration to learn a complex concept).

SRL as Promoting Biology Learning

SRL may advance the learning of science and biology in several ways. It may promote students' cognitive flexibility by enabling them to examine problematic issues based on the structure of their own existing bodies of knowledge and based on their current understanding of knowledge from different points of view and at different moments in time (Spiro, Feltovich, Jacobson, & Coulson, 1991). Engagement in knowledge construction was shown to promote understanding and achievement and to decrease student misconceptions due to learners' increased ability to make meaning of the newly acquired information at their own pace and according to their own existing knowledge. Experiencing autonomous learning enables students to practice different cognitive skills that enhance information processing of complex biology content (e.g., identifying causal relations, shifting between levels of phenomena, or interpreting visual representations).

Self-regulated learners may approach topics in a sequence that considers their own knowledge, understanding, and abilities; thus, such students may shift back and forth or sidetrack while dealing with the different content in an order that makes sense to them. Students may also learn at their own preferred pace, may organize their learning environment according to their individual needs, or may decide to acquire a skill they lack to improve task performance when solving a problem.

The SRL mode may also decrease the reported difficulties inherent in the visual representations that are commonly employed to promote learning of complex scientific concepts and processes. Determining one's own pace enables learners to utilize their time to interpret representations, increase comprehension by mapping between representations and the relevant written text, and construct a coherent body of knowledge by integrating the

information from different visual or textual sources. In addition, self-regulated students can choose a familiar or preferred visual representation to complement textual information. Eilam (2012c) reported that often visual representations used in the science classroom or even in textbooks do not gain students' and teachers' attention and usually go unnoticed. While learning alone, students may invest time and effort into making sense of these representations as a tool for supplementing the studied textual information. Hence, SRL may support student attempts to overcome the difficulties inherent in learning biology.

Self-Regulation of Science Learning and Field Study Interventions

Self-regulated learners construct knowledge and understanding by becoming actively involved in the learning process, for example, by linking facts with explanations (DeLeeuw & Chi, 2003). Metacognitive abilities like monitoring and self-reflection may promote students' mindful choices of strategies for achieving set goals, thereby improving learning processes and outcomes (Brown & Pressley, 1994). These processes may be further advanced by different prompts and teacher guidance. For example, a study exploring high school students' SRL in physics reported that those involved in teacher-guided practice of self-regulation used more strategic behaviors and for longer time periods than those who were regulating learning spontaneously without prompts. Moreover, the knowledge gains of both self-regulated student groups were significantly higher than those of students taught traditionally (Manlove, Lazonder, & de Jong, 2007).

SRL in science has been studied in different contexts (Azevedo & Cromley, 2004; Eilam & Aharon, 2003; Eilam et al., 2009; Peters & Kitsantas, 2010; Schraw et al., 2006). For instance, Azevedo and Cromley (2004) examined university students' SRL while studying the human circulatory system using hypermedia. Findings showed improvement in regulation skills and in knowledge construction. In another study utilizing a computer-based learning environment, Lee et al. (2009) investigated college students' comprehension of the human heart as a biological system as well as their self-regulation, after practicing active, generative learning with metacognitive feedback using a computer. These students performed significantly better on a comprehension test and applied more regulation strategies than students in a control group. In addition, a field study examined junior high school students' SRL behaviors in the domain of ecology and their correlations with student personality traits (Eilam & Aharon, 2003; Eilam et al., 2009). Data were collected via routine protocols completed by students in each lesson, which enabled students to compare their planned and enacted activities, thereby producing cues for improving monitoring. Findings showed improvement in students' SRL behaviors along the year, and SRL significantly correlated with academic achievement and with conscientiousness.

To promote eighth graders' science knowledge in the course of an inquiry, Peters and Kitsantas (2010) used embedded metacognitive prompts while applying Zimmerman's (2000) four developmental levels. They contended that "students who use explicit metacognitive strategies can evaluate their thinking to determine if it aligns with the rigorous requirements of science" (p. 384). Both experimental and control groups gained significantly on pre-post measures of metacognition, content, and science knowledge, probably due to the constructivist nature of the intervention. However, no differences emerged between the two groups' gains, which the authors attributed to students' age and their inability to explicate their thinking. Last, Cleary and Labuhn (2013) applied Zimmerman's (2000) triadic model (as described below) to develop the Self-Regulation Empowerment Program field intervention in science. They found that this intervention promoted high school students' SRL and science achievement. These and other studies corroborated the importance of providing earlier opportunities in school for practicing and fine-tuning SRL

processes, in particular while learning science (Greene & Azevedo, 2007; Kapa, 2007; Lin, 2001).

SRL in the Present Study

We view SRL as conceptualized in Zimmerman's (2000) triadic model that includes three phases of SRL: forethought, performance, and self-reflection. Inasmuch as this model conceives SRL as a process, the model enables analysis of student-applied SRL over time. It is also highly applicable in authentic school learning situations (Cleary & Labuhn, 2013). According to Zimmerman's view, SRL is perceived as the product of multiple interactions among personal, behavioral, and environmental factors. It is an iterative phenomenon involving three types of dynamic processes. These triadic processes constitute feedback that the learner perceives based on past performance and that bring about the adjustment of current performance. This perceived feedback involves the adjustment of self-regulation regarding strategic performance and self-observation, environmental conditions, and cognitive and affective states. Within this view, Zimmerman's applied model includes the three phases of forethought, performance, and self-reflection, as described next.

The Forethought Phase. This phase includes task analysis and corresponding goal setting and strategic plan design. A change in task necessitates changes of goals and plans and therefore requires continuous adjustment. Beyond goal orientation, the forethought phase also includes self-motivation beliefs, namely, self-efficacy beliefs about one's ability to perform the task at hand, as well as internal interest. Locke and Latham (2006) developed the goal-setting theory that focuses on specific goals; accordingly, high-level goals lead to greater efforts and persistence and thus also to a higher level of task performance than low-level goals. Locke and Latham (2006) showed that setting goals is a discrepancy-creating process, a notion applied in the design of our current instruments. Within goal-setting theory, goal orientation/achievement theories explain achievement behavior as related to performing an academic task in school and as concerned with the reasons why individuals desire to perform the task correctly and how they approach and perceive the task. Goals reflect a standard for judging performance, which in turn affects future behavior (Pintrich & Schunk, 1996).

Researchers have identified two kinds of goal achievement theories: mastery goal theories and performance goal theories (Pintrich, 2000). Elliot (1999) added performance avoidance theory. Mastery goals, which focus on task mastering and the development of competence and therefore contribute to meaningful learning, are preferred over performance goals. The latter indicate intentions to learn superficially and to demonstrate competence relative to others or to an external standard, due to the desire to complete the task. Elliot contended that individuals may also be motivated to act due to their desire to avoid demonstrating a lack of competence. Research has suggested that low confidence in one's capabilities is associated with goal avoidance motivation, which in turn results in lower goals and poorer performance. In contrast, autonomy goals evolving from intrinsic motivation lead to mastery goals and better performance, similarly to approach goals (Locke & Latham, 2006). It is suggested that individuals should adopt an orientation suitable to their ability level and the situation (Bell & Kozlowski, 2002).

In all, the forethought phase includes three main factors: self-efficacy beliefs, internal interest, and goal orientation. Together, these support students in sustaining the continuous investment of efforts needed to learn, despite difficulties or obstacles.

The Performance Phase. In this phase, students describe how to proceed while executing the task, organize their environment to assure that it will not distract them from completing the task, and apply the strategies selected for enhancing performance. During this phase, students apply self-observation to track their advancement on the task and may use self-recording to trace specific elements of their performance and its outcomes as related to the set goal and strategies applied, namely, on task monitoring.

The Self-Reflection Phase. This third phase occurs after task completion. In this phase learners evaluate their performance in light of evidence and the outcomes, and they attribute causes for their learning products. Self-evaluation evolves from ongoing monitoring processes and from perceiving the advantages of the strategies applied to the specific conditions or the disadvantages of those strategies that impeded performance. The completion of such a feedback loop constitutes the basis for adjusting performance the next time.

The Triadic Model's Current Application. There are many ways in which this model may be applied in the field, as expressed in the different learning environments designed and the various tools used previously to enact and measure SRL. The present, novel intervention differed from other interventions in several fundamental ways. First, it was a long-term, yearlong intervention, enabling students to experience many phases of regulation and monitoring or even failure along the year, but also many opportunities to sustain their motivation despite obstacles and to adjust their performance. This temporal dimension also enabled online investigation of SRL and the complexity of its processes. The long-term design aimed to contribute to current knowledge about SRL and its development over time in this specially designed environment and about its relation to student development of domain knowledge in biology (Azevedo, 2009; Eilam, 2012b).

Second, because we aimed to examine the benefits of applying SRL to biology learning in typical classrooms, in the current intervention we used the national biology curriculum as studied in a heterogeneous classroom of a typical public school, rather than developing special curricular materials for this study or selecting a specific sample of participants. Third, we used specially developed instruments to promote students' SRL and to capture changes over time in students' adjustment of performance. In the next section, we describe these instruments and their various components as responding to the need to capture students' SRL.

SELF-REPORTING INSTRUMENTS AND PROMPTS

SRL may be scaffolded by prompts. Prompts are hints that may be used for initiating productive deep learning processes, and they are presented to learners on paper or on a computer (Berthold, Nuckles, & Renkl, 2007; Eilam & Aharon, 2003; Zimmerman, 2008). They may activate individuals' existing strategies (Berthold et al., 2007), focus attention on important ideas or core aspects of a task (Ge, Chen, & Davis, 2005), or initiate task-related cues that involve relevant prior knowledge (thus also initiating monitoring that is directly relevant to the task at hand, which promotes performance) (Kapa, 2007). Higher effectiveness of prompts emerged when students were forced to respond to them rather than skip them (Ge et al., 2005). This was especially true for situations involving performance of tasks in an open-ended environment, which imposed higher demands on students' cognitive and metacognitive abilities (Bell, Davis, & Linn, 1995; Ge et al., 2005). Findings suggest that prompts have differential effects on students regarding the desired outcomes (e.g., Chi, DeLeeuw, Chiu, & LaVancher, 1994; Ge et al., 2005; Peters &

Kitsantas, 2010). For example, researchers reported that self-monitoring prompts showed higher benefits for students who identified difficulties in their tasks than for students who perceived no difficulties (Davis & Linn, 2000). Greene and Land (2000) suggested that prompts alone are insufficient because students do not spontaneously engage in processing. A German study demonstrated the influences of protocols containing prompts on students' SRL and its products (Nückles, Hübner, & Renkl, 2009). The students in that study's experimental group (who completed the prompted protocols) were more focused on the learning assignment, planned their activities better, showed better monitoring and regulation of learning, and understood the studied content more than their control counterparts who freely completed unprompted protocols.

Most self-report questionnaires measure SRL as an aptitude or trait across situations and tasks, requiring students to search their memory and make an accurate generalization about their SRL (Braten & Samuelstuen, 2007). For example, the Learning and Study Strategies Inventory (LASSI; Weinstein, Palmer, & Schulte, 1987) taps students' perceived self-reported SRL activities applied generally while learning, with a focus on thoughts and behaviors related to successful learning. However, if SRL is perceived as a process as conceptualized in Zimmerman's (2000) model, then an online measure is the optimal approach, involving think-aloud, observation, or trace instruments (Veenman, 2011). This online "process" approach was adopted for developing the three protocols utilized in our current study: the Weekly Self-Report Instrument (WSRI), Yearly Self-Report Instrument (YSRI; Eilam, 2002), and Test Self-Report Instrument (TSRI; developed by the second author for the current study). Although these three instruments are self-reports, they differ substantially from self-report questionnaires like LASSI and more closely resemble think-aloud protocols that include metacognitive prompts. In completing the WSRI, YSRI, and TSRI, students did not need to search their memory to generalize across different tasks and recall information, rather they reported online about their activities and goals in the context of the specific task and immediately after its enactment. These instruments also enabled students to evaluate their own progress in learning.

Specifically, the WSRI aimed to scaffold learners' metacognition as they planned, executed, monitored, reflected on, evaluated, and adjusted these activities during their weekly science lessons. Perceived cues regarding differences between the suggested and enacted plans could initiate self-evaluation processes regarding the efficacy of the planning process and the changes needed in light of the specific conditions encountered. Self-evaluative judgments are linked to causal attributions about outcomes, and they depend on cognitive appraisal of personal and mitigating environmental conditions (Zimmerman, 2000). The YSRI was based on the same principles as the WSRI but presented learners with a yearlong scale to scaffold and support their long-term learning processes, aiming to enable students to evaluate their progress on their yearlong task and to perform the necessary adjustments. The TSRI was based on the same principles but related specifically to science test taking. By raising students' awareness to cues about the quality of the knowledge they acquired in preparation for the science test, compared to their actual test performance, students could evaluate the specifics of their test preparation processes and their learning along the year. These instruments were all developed to increase objectivity, uniformity, and validity of self-reporting, reduce the effects of cognitive load and social desirability, and document students' SRL and their relevant behavior (see the Method section for detailed descriptions of these tools).

Our paper-and-pencil instruments have common features with the "metacognitive tools" suggested by Azevedo (2007) for computers. For example, these instruments (a) support cognitive processes for accomplishing the task such as setting goals, sequencing activities, organizing resources, and examining the capacity of the context (e.g., setting, objects,

peers, teachers); (b) decrease students' cognitive load by externalizing information on paper/computer, thus freeing mental efforts for high-order processing; (c) support students in sustaining motivation while performing complex long-term learning by enabling ordered documentation of the process and examination of past experiences; and (d) prompt cognitive and metacognitive activities that may advance learning.

Our instruments—protocols—required learners' continuous online reporting about their performance to “calibrate” their behaviors. Calibration is metacognitive monitoring—a component skill of SRL referring to individuals' accuracy of online perceptions and judgment of their own cognitive performance as related to the set goals and a criterion task. It characterizes how aware students are of what they do/do not do and know, which is a necessary precondition for successful learning (Pieschl, 2009). In the present case, students calibrated their performance after completing the enactment of plans, as well as after receiving their test back from the teacher, and also along the year. Calibration was the basis for the next iterative phases of SRL—made possible by comparing suggested versus enacted plans in all three instruments and by comparing enacted plans' outcomes versus the set goals. The discrepancies perceived from such comparisons (which sometimes were implicit) acted as feedback cues regarding the need to calibrate planning and actions (e.g., by changing the time allocated to different activities, switching activities' sequence, deciding to apply more efficient strategies in future performance in similar contexts, altering goals or even abandoning them).

Different factors have been reported as influencing the accuracy of calibration. For example, in Nelson and Dunlosky's (1991) examination of the accuracy at which undergraduates calibrated their knowledge while performing a simple well-defined task (learning 66 paired associates followed by a recall test), accuracy was higher for delayed judgments of learning than for immediate judgments. Lin and Zabrocky's (1998) review indicated that the effect of text difficulty on calibration has not yet been determined in the case of reading, whereas active processing usually was found to increase calibration accuracy. To account for such influences, in the current study, both immediate and delayed judgments were applied (e.g., immediately after the enactment of each activity and after the enactment of the whole plan at the end of each lesson).

We next elaborate on the components of these protocols, as reflecting Zimmerman's model of the three iterative phases of SRL: forethought, performance, and self-reflection.

Planning. Planning is an intentional process aiming to ensure successful task performance and the achievement of a particular goal under specific circumstances (Abrahams & Reiss, 2012; Prins, 2002). It is part of the forethought phase, which is the first phase of Zimmerman's (2000) model, namely, the initial understanding of the task and goals before acting on them. Experiences with a less structured environment have been shown to result in greater planning than in structured settings (Jordan, Ruibal-Villasenor, Hmelo-Silver, & Etkina, 2011), endorsing the open-ended learning environment in our study as appropriate for promoting SRL. Planning in an open-ended dynamic learning environment is particularly difficult because it requires coping with unknown variables and is characterized by uncertainty about actions' results (Allen, Hendler, & Tate, 1990), as in the present study.

Planning allows for anticipation of outcomes, prevents some mistakes, and involves the transformation of learning intentions (goals) into action plans (Gollwitzer, 1996). After setting the goals following the task's analysis, the high-level abstract distant goal is decomposed into limited proximal attainable goals, which increase the potential for goals' achievement (Manlove et al., 2007; Zimmerman, 2000) in each of the weekly learning sessions. The WSRI promoted students' ability to initialize goal-relevant actions by

selecting a set of activities in light of available resources and determining their enactment sequence. The resulting generated external representation (i.e., on paper) of individual's future behavior prior to enactment may guide future planning in similar situations, hence improving and refining SRL in the following feedback (Prins, 2002; Zimmerman, 2000).

Monitoring. Enactment of plans constitutes the second phase of the model, namely, performance. Monitoring is part of the performance phase in Zimmerman's (2000) model. If execution of plans is not monitored—to evaluate and revise them accordingly—then no change in behavior would be possible. In the current study, learners monitored their enacted activities in terms of these actions' capacity to advance learners toward their learning goals. The standards for successful monitoring are one's plans, set goals, environment, and outcomes of enactment as related to the goals. Perceived monitoring feedback enables the refinement of plans and behaviors for future use in similar circumstances (Schraw et al., 2006)—in our case based on Zimmerman's model—a new weekly cycle. Consequently, monitoring may lead learners to adjust their initial plans by introducing, modifying, and/or omitting certain activities in the course of task performance or even by changing initial goals (Prins, 2002).

While monitoring progress toward their stated goals using the YSRI and WSRI, students could perceive cues initiated by gaps between their current situation regarding the task at hand and their desired situation (Papaleontiou-Louca, 2003; Zimmerman, 2002). On the basis of Butler and Winne (1995) study, we defined cues as signals, metacognitively perceived by individuals, regarding variance between current and expected states, which in turn stimulated them to react accordingly. In our study, the WSRI and YSRI were designed to initiate such perceptual cues, which have been found to be more effective than external cues provided by others such as teachers or peers (Ryan & Deci, 2000). Perceptual cues may be affective (e.g., "I hate this topic"), personal (e.g., "I am too tired to concentrate on the problem"), environmental (e.g., "It is too noisy for me to learn here"), or cognitive in nature (e.g., "I still don't understand this process"). Based on perceived cues, students can reevaluate their plans in light of goals and find solutions for identified problems in learning (Butler & Winne, 1995).

Time Management During Planning and Monitoring

Although students' perception of time may determine how they engage with tasks (Duncheon & Tierney, 2013) and although time management is a core component of planning and monitoring processes (Eilam & Aharon, 2003), very little has been published on this issue as related to science learning. During a long-term autonomous learning process, time constraints in light of task requirements call for making decisions concerning choices among alternative actions. Thus, learners' perception of the rate of their progress toward the goal may affect their planning and their enactment of plans. A progress rate differing from expectations has been found to affect the type of regulation that individuals may perform and their emotions toward the perceived pace, which in turn may influence their planning (Carver & Scheier, 1990). As suggested in Zimmerman's (2000) model, time management is influenced by behavioral factors (i.e., efforts to self-observe, self-evaluate, and self-react to academic performance), environmental factors (i.e., the use of planning prompts), and personal learning factors (e.g., metacognitive knowledge and skills, goal setting, self-efficacy; Zimmerman, Greenberg, & Weinstein, 1994).

THE LEARNING CONTEXT

Teachers' perspectives regarding learning-teaching processes may be viewed along a continuum ranging from teacher-regulated instruction to students' autonomous learning. Each environment has its own advantages and challenges, depending on teachers' perspectives and beliefs, students' characteristics, and tasks' requirements. In line with the literature reporting the benefits of autonomous learning environments for developing SRL, we aimed to promote the desired shift from TC learning to SRL in an urban public school in the center of Israel. We designed a self-regulating (SR) biology class environment that would support and promote such a shift, requiring students to redefine and restructure their knowledge about learning and what it entailed. A second class, the TC class, studied the same content via a teacher-regulated mode, serving as a control group.

The Two Study Groups

In the SR classroom, students managed their own learning assignments and modes. This environment required self-determination and presented students with many challenges. They acted according to interests, capabilities, and beliefs, whereas teachers' guidance was directed toward scaffolding students' ability to learn. The lack of teachers' instructions aimed to propel SR students to become actively involved in learning and to take responsibility for it. This increased complexity required investment of efforts. Students constructed knowledge and understanding by reasoning, made meaning of new knowledge, and applied acquired knowledge in new situations. They had to constantly self-evaluate their actions and the quality of their products against canonical knowledge, instead of just following teachers' evaluations (Nicol & Macfarlane-Dick, 2006; Perkins, 1998). In addition, in contrast to the TC students' automatic progression according to the linear presentation of contents in the textbook, learners in this SR environment utilized the potential inherent in that setting. They selected their personally preferred order of topics, in accordance with their knowledge structures and understanding, and at their own preferred pace (e.g., allocating time according to needs).

In the TC environment, students could apply SRL spontaneously when they had the opportunity to do so (e.g., while performing homework or studying for a test), but these opportunities were limited due to teacher-regulated learning-teaching processes. In contrast, in the SR environment students' opportunities for SRL were unlimited and were supported and advanced by the specially designed instruments. For example, self-assignment of homework reflected SR students' response to an evolving need in the learning process, rather than reflecting a routine task as for the TC students.

The Curriculum

Students in both the SR and TC groups studied the ninth-grade science curriculum, which included three main topics in genetics of increasing complexity: the cell structure including its nucleus, Mendel's laws, and basic genetic-related statistics. These topics challenged students in several ways: The cells, organelles, and molecules are 3D but are presented on a 2D space; these items cannot be seen by the naked eye but only through different visual representations (e.g., pictures, microscope, models); the understanding of gene transfer across generations involves interdisciplinary integration of both biology and statistics; cognitive flexibility is required to learn these new concepts, which may hold different meanings and definitions as learning progresses (e.g., the concept of gene can be examined as an inherited factor composing the chromosome or a DNA unit that carries information); some topics are concrete, and other are abstract; and more. Many studies have reported

the difficulties inherent in learning these topics when studying with teachers (e.g., Lewis, Leach, & Wood-Robinson, 2000; Marbach-Ad & Stavy, 2000); hence, studying them alone must be even more challenging. We expected that SRL would promote students' ability to overcome such difficulties.

RESEARCH QUESTIONS

The present study examined the influence of SRL, as applied by ninth-grade students in a specially designed SR learning environment, on these students' science knowledge (genetics) as compared with students in a TC environment. In addition, SRL was measured by the LASSI in both groups and by three specially developed instruments in the SR group. The latter three protocols were designed with a dual purpose to promote students' SRL and to enable us to follow SR students' activities. Using these three special instruments in the TC environment would be useless due to the teacher's continuous external regulation of students' learning behaviors. Even homework was regulated in the sense that it was administered without evaluating each student's understanding. Therefore, we did not apply a full design but still could infer about the SR students' enacted SRL in the course of learning science. Thus, we asked three research questions:

1. What changes are demonstrated over time by the two student groups in perceived self-reported SRL, as measured by the LASSI for both student groups, and also by the WSRI, YSRI, and TSRI for the SR group?
2. What changes are demonstrated over time by the two student groups in science (genetics) knowledge?
3. What correlations emerge between perceived self-reported SRL and students' science knowledge in both groups (and perceived enacted SRL in the SR group)?

METHOD

Sample

Students ($N = 52$; age 15–16 years) attended two ninth-grade science classrooms that were each highly heterogeneous regarding both students' socioeconomic status and their average academic achievement in the previous grade, ranging from 52 to 95 of 100. No significant differences between the two classrooms were found based on science (biology) knowledge test scores at the end of the eighth grade. One of the classrooms was assigned to learn in the SR environment ($n = 25$; 13 boys, 12 girls), and the other group was assigned to learn in the TC environment ($n = 27$; 14 boys, 13 girls).

Data Collection

Two measures were administered to the entire sample (SR and TC groups) for collecting data on students' SRL behaviors and on their acquired science knowledge: (a) data regarding students' perceived self-reported SRL activities applied generally while learning were collected by LASSI and (b) data regarding students' science knowledge was collected by specially developed science tests. In addition, for the SR group only, data were collected regarding students' enacted SRL in different time frames and contexts—over the whole year, each week, and regarding self-evaluation of test performance—collected by the specially developed YSRI, WSRI, and TSRI, respectively. Next, we describe each tool in detail. Observations of the two groups were also recorded in writing to ensure group fidelity and for the SR group to ensure instrument fidelity and yield supplementary data on self-reports.

LASSI. The entire sample completed the 76-item LASSI (Weinstein et al., 1987) to measure perceived self-reported SRL, focusing on thoughts and behaviors related to successful learning (Cronbach $\alpha = .91$). Students rated themselves on SRL-related statements (e.g., “I do not care about getting general education; I just want to get a good job” or “I only study the subjects I like”) using a 5-point Likert scale ranging from *Not at all like me* (1) to *Very much like me* (5). The Cronbach alphas for the 10 LASSI scales were attitude (.75), motivation (.84), time management (.89), anxiety (.89), concentration (.86), information processing (.87), select main ideas (.76), study aids (.73), self-testing (.79), and test strategy (.83). The LASSI is commonly used to investigate self-reported SRL, and its validity has been well established in diverse contexts and populations. This measure was also shown to predict performance (Cano, 2006).

Enacted SRL. Enacted SRL scores constituted the sum scores of the three protocol types: YSRI, WSRI, and TSRI. These protocols were completed by the SR group only, through the academic year, and provided data (scores) about the SR students’ learning activities during the forethought phase, the performance phase, and the self-reflection phase, as well as the behavioral changes applied in the next SRL cycle. All these protocols were based on students’ perception of cues and calibration.

YSRI. Grounded in the aforementioned literature, this instrument (see Table 1) was developed by Eilam (unpublished) and successfully applied in several studies (e.g., Eilam, 2012a; Eilam & Aharon, 2003). The YSRI was found in previous studies to be essential for supporting students’ time management while learning self-regulating throughout a full academic year, something they had never experienced before. It highlighted SR students’ actual progress in learning throughout the year, compared to the teacher’s suggested work plan. The teacher initially divided the overall proximal goal for the year (learning and understanding genetics) into distinct goals coinciding with the topics to be learned each week (e.g., blood groups or Mendel’s laws). The SR students received a printed form presenting 22 lines, one line for each week. The seven columns represented: date of the week, teacher’s suggested topics for learning, students’ enacted learning, size of gap between suggested and enacted learning (i.e., number of lessons behind or ahead of the suggested plan), reasons for the gap, suggested actions for decreasing gaps, and evaluation of that week’s performance. The YSRI was completed at the end of each week. As suggested by Locke and Latham (2006) in their goal-setting theory, any discrepancy between the suggested yearlong scale and the enacted session stage was expected to initiate students’ metacognitive awareness of their progress on the yearlong schedule and to promote behavioral changes in planning the upcoming activities and their sequence, time allocations, and settings. The YSRI also indicated the dates of tests, enabling students to monitor progress vis-à-vis the teacher’s suggested contents to be learned for each test.

WSRI. This instrument (see Table 2) was also developed by Eilam (unpublished) based on the aforementioned literature and was successfully applied in several studies (e.g., Eilam, 2012a; Eilam & Aharon, 2003). It highlighted SR students’ actual progress during each specific session in a certain week, compared to their detailed plans for that lesson (students had two lessons per week totaling 3 hours—one single hour and one double hour). The instrument comprised four main parts: goal phrasing, two columns of planning and enactment, and self-evaluation. In addition, students were provided with a pool of seven suggested basic activities for use while describing their planned and enacted activities (e.g.,

TABLE 1
The YSRI: Yearly Self-Report Instrument

Week	Teacher's Suggested Plan	Enacted Plan	Size of Gap	Reasons for Gap	Actions That Should be Taken in Light of Gap	Self-Evaluation of Performance
1/9	Genetic inheritance—what is that? Cell structure (Chapter 1)					
8/9	Cell structure and nucleus function					
15/9	Acquired vs. inherited characteristics (Chapter 2).					
22/9	Chromosomes, genes, DNA and traits (Chapter 3).					
29/9	Review, Chapters 1–3. Test 1: Inheritance throughout generations.					
.
.
.

TABLE 2
The WSRI: Weekly Self-Report Instrument

Date: _____

My learning goal this week: _____

Time	Planning		Enactment		Self-Evaluation of Performance	
	Activity	Time Allocation	Activity	Time allocation	Score (1 to 10)	Possible Reasons
10:30						
10:40						
10:50						
11:00						
11:10						
11:20						
11:30						
.						
.						
.						
.						

Homework: _____

Weekly feedback:

Is there a gap between planning and execution? Yes/no (circle the right answer).

Where is the gap? Time / type of activity / order of activity (circle the right answer—there can be more than one factor. If there isn't a gap, erase all the possibilities).

Did I reach the weekly goal? Yes/no (circle the right answer).

Did I define the weekly goal well? Yes/no (circle the right answer).

What are the ways or actions I have to take in order to reach my weekly goal? Elaborate: _____

answering textbook questions, summarizing content, drawing a concept map). Students could add activities to this list as needed.

At the beginning of each lesson, students phrased their specific short-term weekly *goals*. They could base their selection on the teacher's suggested scale or the textbook's linear presentation of the topics, or they could suggest goals according to their own understanding. This focused students on limited achievable session goals that would facilitate monitoring (Hacker, 1998) and thus sustain self-efficacy while engaging with the long-term task (Schunk, 1990).

Then they outlined their detailed plans for that lesson in the "planning" column, by indicating the *planned activities* to be carried out to achieve the stated goals, ordering these activities' exact *sequence* of enactment, and allocating *time* for each activity. The self-evaluation part in that planning column constituted two questions (not shown in the table): (1) Will the plan enable the achievement of the weekly goals? and (2) Will you be able to enact this plan? Students had to respond to each of these questions on a 5-point Likert scale, ranging from *Not at all* (1) to *Absolutely yes* (5). Both of these self-evaluations derived from students' past experiences with different planned activities and therefore were assumed to reflect students' self-efficacy. After completing their initial plans and self-evaluations, students began enacting their plans and adjusting those plans as necessary to the situation at hand. After completing each activity, they used the "enactment" column of the weekly protocols to record in sequence their *enacted activity* types, and their exact time duration.

Perceptual cues, initiated by comparing the planned and enacted activities, were expected to promote students' ability to monitor progress; aiming to achieve the set goal, they could identify gaps between planned and enacted activity types, gaps in activities' time duration, and gaps in order of enactment. Students' awareness of gaps and of the relations among plans, enactment, and outcomes promoted their ability to generate a more effective plan in the next cycle performed in the coming lessons. Acknowledgment of the gaps increased the search for the causes of these gaps, based on the differences between students' own past experiences and the present conditions. For example, students could ascertain that although they could usually answer their textbook questions in 30 minutes (thus, while planning, they had allocated 30 minutes for this activity), this week the questions' enactment took 55 minutes because they faced some difficulties in the second question that required rereading of the chapter. Following the enactment of each activity students performed a self-evaluation that they recorded in the "enactment" column of their WSRI. They evaluated their learning in each activity on a 10-point Likert scale, ranging from *Did not learn at all* (1) to *Know it perfectly* (5). Finally, given their plans and modes of enactment, students indicated the extent to which they achieved the weekly goal, pointed out possible causes that inhibited or contributed to their success, and suggested actions that would advance them closer toward the set goals. Among such actions, students could regulate their comprehension, skill efficiency, and progress by assigning themselves homework.

The WSRI self-report resembled prior "metacognitive tools" (Azevedo, 2007) in some ways (e.g., as a self-report, as continuously completed by learners in real time over the yearlong task performance, or as requiring students to reason in writing about their performance to calibrate their behaviors). However, it also differed from such instruments in several key features that increased its validity and achieved additional benefits:

- a. *The use of given activity segments to increase accuracy of reporting as well as objectivity and uniformity in self-reporting.* In completing the WPSI in each session, students could either choose an activity segment from the given pool of seven potential learning activities, phrased in familiar school-related language (e.g., "define

a concept”), or else could suggest new activities using their own words. Hence, students did not have to label their activities according to an academic register. The chances for subjective errors in such descriptions were much smaller than when students assessed their perceived SRL performance using other self-report instruments like LASSI, where students’ unawareness of an automated strategy that they used or their failure of memory might lead to misreporting. In addition, the given pool of segments yielded a uniform “language” of reports, which facilitated student comparisons between the planned and the enacted parts within a session as well as comparisons of reports between sessions, thereby enabling cue input and monitoring of plan effectiveness. For example, it enabled students’ easier identification of repetitive sequences of efficient activities, which could bring about the refinement of the next SRL cycle. Finally, the segments’ uniformity and objectivity minimized room for researcher interpretation by increasing reliability (interjudge agreement) when coding the segments inserted in different WSRI.

- b. *Reporting of enacted activities immediately following their completion, rather than at the end of the week, thus reducing the effect of cognitive load and recall errors.* Reporting and monitoring of progress straightaway after performing a meaningful activity that took some time has been found to decrease cognitive load (Winne, 1995). This online mode of reporting activities has greater validity and reliability than offline measures because students do not need to search their memory for past activities and make generalizations regarding how much or how often they applied a specific strategy, but rather they report their current task- and context-specific strategies as actually enacted (Bannert & Mengelkamp, 2007; Bråten & Samuelstuen, 2007).
- c. *Legitimacy of reporting any enacted activity content, sequence, and timing, thus reducing social desirability.* In the open, autonomy-based learning environment, any student’s suggested or enacted activity, explanation, reason, etc. were legitimate. No penalties of any kind were imposed on an off-task behavior, excluding disturbing peers’ learning.
- d. *After reporting activities for the whole session, students could evaluate the effectiveness of that session’s activities in achieving the session’s set goal, thus minimizing social desirability.* The clear session goal written at the top of each WSRI served as an anchor for comparison of planned versus enacted activities, permitting students to seek their own success in meeting their set goal rather than seeking competition with other students’ goals or satisfying the teacher’s expectations.

WSRI Validity. The validity of the WSRI was measured in line with Veenman’s (2007) indices for assessing the validity of new measures, as follows: First, a high interrater reliability of WSRI activity segments emerged (90%) in a prior study that examined 20% of students’ SRL reports (Aharon, 2003). In the present study, coding resulted in close to 90% interrater agreement. Second, the WSRI demonstrated convergent validity indicated by a significant positive correlation between enacted SRL measured by this instrument and SRL measured by the LASSI, both in a prior study, $r = .45, p < .01$ (Aharon, 2003), and in the present study, $r = .52, p < .05$. Third, external validity was found in the aforementioned study, where regression findings showed a significant moderate correlation with the enacted SRL measure, $r = .35, p < .001$, contributing more to the prediction of achievement ($\beta = .39, p < .01$) than did the LASSI self-report measure ($\beta = .31, p < .05$). In the present study, enacted SRL was also found to be a significant predictor of science achievement scores, accounting for 26% of the variance ($\beta = .51, p < .05$).

TSRI. This protocol elicited internal cues regarding SR students' descriptions of the learning activities that they applied while studying for tests and descriptions of their conduct during test time. The aim of the TSRI was to help students improve self-regulation of their test preparation process as well as their test performance. As part of the TSRI, while taking each science test (at five intervals—see description below), students wrote their expected grade based on their self-evaluation of their current knowledge level on the tested contents. After receiving the graded test back from the teacher (with errors marked), the SR students completed the TSRI, where they were required to (a) identify gaps between the teacher's grade and their expected grade; (b) record their number of incorrect responses for each science knowledge type in the test (declarative knowledge, understanding, and application ability); (c) indicate the type of knowledge they should strengthen; and (d) suggest reasons for the gaps and activities that should be undertaken, before and during test, to improve future performance and learning outcomes.

Science Knowledge Tests. The entire sample completed science knowledge tests at five intervals, designed to assess students' (a) declarative knowledge, (b) understanding, and (c) knowledge application ability regarding the three terms' different curricular contents. The teacher developed three 25-item science knowledge tests, one for each of the three terms. The first test (after Term 1) provided a baseline score for student knowledge of the previous year's biology contents, comprising cell structure and the nucleus, as related to heredity and to acquired traits. The second test (administered before and after learning in Term 2) assessed student knowledge of Mendel's laws and intergenerational traits' transfer according to these laws. The third test (administered before and after learning in Term 3) assessed student knowledge of the heredity laws of codominancy, blood groups, and quantitative statistics of heredity.

Each test's 25 items included three parts. Part A contained nine closed multiple-choice questions assessing *factual, declarative knowledge* (i.e., mainly recall and memorization), tapping knowledge of facts and definitions regarding concepts and processes (e.g., "The definition for chromosomes is . . ."). Part B contained 11 open-ended questions, of which eight assessed *understanding and near application*, demonstrating students' ability to explain concepts and phenomena in their own words as well as to solve simple problems (e.g., "A corn cell has 20 chromosomes. How many chromosomes are there in a corn sex cell?"). The other three questions in Part B examined students' ability to *apply acquired knowledge* in unfamiliar situations that were not previously presented or discussed in class—*far application* (e.g., "If all living creatures have the same nucleotides, what will happen if we transfer DNA string from a pig to a mouse?"). Part C presented to students an unfamiliar abbreviated scientific paper followed by five open-ended questions. One of the questions required students to identify specific *factual* information in the paper. Another two questions required students to express their *understanding* of certain phenomena in the paper, and two other questions required students to *apply* the knowledge they acquired during the term to the novel contexts described in the paper.

Two expert biology teachers validated the tests' questions by checking their clarity, their relevancy to the content of each term, and their ability to evaluate knowledge, understanding, and application of that content (construct validity). Lack of content or construct validity resulted in changing or dropping items and designing new ones.

Observation Notes. Written notes regarding students' actions in the classroom space were recorded in both groups by the first author five times along the year. These observations were conducted to assess the teacher's consistency in preserving the two different learning

environments, and in the case of the SR group also to ensure fidelity of the instruments and to improve our understanding of student descriptions in the self-reports. Although our focus in the present paper is quantitative, we bring some examples of these notes to enrich the meaning beyond the numbers.

Learning Environment and Procedure

The yearlong intervention coincided with the junior high school's three-term academic year (5-week fall term, 9-week winter term, and 8-week spring term). Both the SR and TC groups studied biology for 3 weekly hours (2 consecutive hours and a separate single hour), in the same fully equipped school laboratory. Students in both classes studied the same standard Israeli genetics curriculum for the ninth grade and were taught by the same expert biology teacher (with 26 years of experience), who was highly committed to enacting the established instructional mode in each classroom (SR and TC). Both groups used the same textbook, followed the same timeline of contents in each term, and were assessed for science knowledge at the same time points by the same set of five tests over the year.

SR Group. Each of the students in the SR classroom made their own decisions regarding all aspects of their own learning of biology. They used the support of the various self-report instruments to guide their selection of goals, to construct and enact plans, to evaluate their own performance, and, based on the cues perceived and their reasoning about them, to adjust and refine behaviors. For instance, students in the SR group could choose to work near/far from the window, alone or near others, and individually or in collaboration with peers. They could assign themselves different tasks at their desired pace and could manage their own time during lessons, and as they gained new experiences over time and became more aware of their personal characteristics as learners, they could rearrange their environment and choices of task, sequence, and time allocation accordingly. They could also alter the sequence of contents studied from that presented in the book or by the teacher, according to their own understanding of these contents. The lab included all the tools and materials required for the performance of the experiments recommended by the curriculum. Students could perform them individually, in dyads, or in small groups, and could be helped by the teacher and the laboratory assistant. The teacher was available for any clarifications or for supporting students, consistently encouraging students to self-regulate their own learning.

TC Group. Students in the TC classroom learned the genetics curriculum in a TC environment. All of them simultaneously experienced teacher-determined modes of learning, contents, homework, pace, settings, procedures, experiments, and demonstrations. They completed the same examinations at the same times as the SR students. After their science knowledge tests were graded by the teacher, students' common mistakes were discussed in the classroom. In this group, too, the teacher was available to support student learning at all times.

Procedure. The LASSI was administered to all students in both groups twice: before the intervention began and after completion of the intervention, to assess any pre-post differences in students' perceived self-reported SRL. In the first weeks of Term 1, students in the SR group received some initial explanations and training from the teacher to guide them in utilizing and completing the WSRI and YSRI protocols appropriately as they began their learning. In the TC group, students began learning in the controlled environment. At

TABLE 3
Criteria for Coding Enacted SRL Recorded in YSRI

Score	Criteria Definition
0	No response
1	Noting “I continued” instead of indicating the actual topics studied that week.
2–3	Clear answer, accurately identifying gap. Identifying the number of hours lagging behind teacher’s suggested schedule.
4	Suggesting possible reasons for identified gaps.
5	Eliciting reason-relevant actions to be taken for closing gaps.
6	Demonstrating delayed behavioral change (e.g., a delayed change in behavior exhibiting perception of cues and relevant actions over time).

the end of the first term, all students performed the first biology test, which was graded by the teacher and returned to students. Only the SR students then completed the TSRI regarding that test. Pretests were administered at the beginning of Terms 2 and 3 to both groups, to evaluate student preexisting knowledge of the relevant term’s contents. Identical posttests were administered to both groups of students at the end of each of these terms to examine science knowledge gains. After each test, only SR students completed the TSRI.

Data Analysis

LASSI. Each student’s score on the LASSI questionnaire was based on the total sum of scores received on LASSI’s different scales items following reverse scoring of appropriate items. To identify differences between the two groups’ perceived self-reported ability to manage their own learning in the TC versus SR environments, pre–post gains were calculated using repeated-measures multivariate analysis of variance (MANOVA) with Cohen’s *d* for effect sizes.

Enacted SRL. Data collected using the completed YSRI, WSRI, and TSRI protocols over the year were analyzed to determine SR students’ enacted SRL. The YSRI data were coded according to three predetermined criteria reflecting awareness of cues and monitoring within this protocol: identifying gaps, reasoning about them, and suggesting actions that should be taken accordingly. As seen in Table 3, for each YSRI criterion, scores ranged from 0 (empty space, reflecting lack of response to the protocol) to 6 (reflecting a behavioral change over time).

Students’ WSRI data were coded according to the four parts of this protocol. First, their phrasing of the goal was coded as either a mastery or a performance goal. The second and the third parts of the protocol were coded according to the protocol suggested and enactment components, and the fourth part of the protocol (i.e., the self-evaluation) was coded to reflect students’ self-evaluation of the previous parts (see Table 4). The criteria and scoring described in Table 4 were applied for each single WSRI protocol and then by comparing weekly protocols over the year to identify and reveal changes in student behavior over time. Overall, WSRI scores could range from 0 (no response) to a total of 21 highest possible score (full correct responses), which is received by adding up all the highest scores in sections A–H in Table 4.

The TSRI items were coded using similar criteria, as seen in Table 5, from 0 (no response) to 6 (changing behaviors over time).

Scores on the three protocols were summed up for each term separately and for the whole year to provide each student’s enacted SRL scores. Mean group scores were calculated for

TABLE 4
Criteria for Coding Students' Enacted SRL as Recorded in WSRI

Criterion		Score	Criteria Definition
A	Goal phrasing	0	No response
		1	Performance goal
		2	Mastery goal
B	Planning/enactment	0	No response
		1	Inaccurate and vague phrasing of activities, which hindered enactment and monitoring (e.g., "Chapter 4," "I studied")
		2–3	Use of external sources in activities (e.g., textbook, teacher-suggested segment pool)
		4	Perception of cues as expressed in adding self-generated activities (e.g., searching additional information via internet, building a model of DNA)
C	Evaluation of plans as related to goal	0	No response
		1	Evaluation of plan's likelihood to lead to achievement of stated weekly goals
		2	Evaluation of goal as indicative of what should be performed in this session (e.g., a general goal "to proceed in the study" hinders planning and monitoring)
D	Evaluation of self-efficacy	0	No response
		1	Any evaluation
E	Identification of temporal, sequential, or planned-enacted activity gaps	0	No response
		1	Identification of about half of existing protocol gaps
		2	Identification of all existing protocol gaps
F	Match between students' evaluation of performance and students' reasoning about it	0	No response
		1	Evaluation does not match students' reasoning (e.g., Performance evaluation = 5 whereas reason indicated = "I understand and know the Mendel laws well")
		2	Evaluation fully matches students' reasoning (e.g., Performance evaluation = 9 and reason indicated = "I understood it well and I compared it with my friends")
G	Suggested activities for improving goal achievement	0	No response
		1	Indication of vague activities that would not improve future plans and goal achievement (e.g., "I will invest more efforts")
		2	Indication of relevant activities that can be enacted to improve future plans and goal achievement (e.g., "I have to complete questions 5 to 7 at home after I ask the teacher to explain the concepts to me")
H	Behavioral changes over time	6	Plans exhibiting a behavioral change indicative of students' monitoring and consideration of past experiences (e.g., a student who reported three times that he did not achieve his goal because he did not understand the content, and then allocated time in his fourth plan for talking with the teacher about "the statistics I did not understand")

TABLE 5
Criteria for Coding Students’ Enacted SRL as Recorded in TSRI

Score	Criteria Definition
0	No response
1	Erroneous identification of gap size between student’s expected grade and teacher’s assigned grade for unknown reasons (e.g., error in calculating gap, lack of attention) but still exhibiting awareness of differences (e.g., expected score 85, teacher score 66, gap of –14 points).
2	Correct identification of gap size between student’s expected grade and teacher’s assigned grade, exhibiting awareness of differences (e.g., expected score 70, teacher score 82, gap of +12 points).
3	Any correct explanation of student’s error on test (marked by teacher) regarding student’s declarative knowledge, understanding, or application of knowledge as related to canonic science knowledge (e.g., “I thought that the name homologous in this definition refers to gender, which is not so; it related to their shape and gene quality and order”).
4	Identification of deficiencies in content knowledge and skills in light of previous explanations and ways to strengthen them (e.g., “I should seek teacher’s help more often” or “I should read about it”).
5	Suggesting relevant reasons for the identified gap (e.g., “I really thought I understood it when I read the book” or “I thought two days before the test were enough” or “I spent too long struggling with one question”).
6	Changing behaviors over time as expressed in protocols’ suggested activities regarding ways to study for a test (e.g., “I have to build a concept map rather than just write concepts”) and ways to perform during tests (e.g., “I have to read the whole test quickly before beginning to answer”).

each term and for the whole year to show group changes in enacted SRL behaviors over time. Reliability between two coders for a random 20% of the sample (coding all three protocols) reached close to 90% (after some rounds of redefining criteria). Differences between enacted SRL scores in each term were calculated using repeated-measures MANOVA and post hoc tests.

Science Knowledge Tests. Each of the 10 questions assessing declarative knowledge (facts and definitions) was scored up to 2 points (20 points in total), each of the 10 questions assessing understanding was scored up to 5 points (50 points in total), and each of the 5 questions assessing students’ application ability in new situations was scored up to 6 points (30 points in total), yielding a total score ranging from 0 to 100 points. As may be seen, scoring placed heavier weight on students’ understanding and application of constructed knowledge than on rote memorization. Students’ science knowledge scores for each term and their average science score for the whole year were summed.

Three experts in biology independently coded a random sample of 20% of student responses to the three tests. Interjudge agreement reached 87% through a process of discussions and criteria refinement. MANOVA, *t* tests, and Cohen’s *d* for effect sizes were used to examine differences between groups’ knowledge at the end of Term 1 (baseline), and between their knowledge in Term 2 (pre–post) and Term 3 (pre–post).

Observation Notes. Notes were used as-is for validating environments and instrument fidelity.

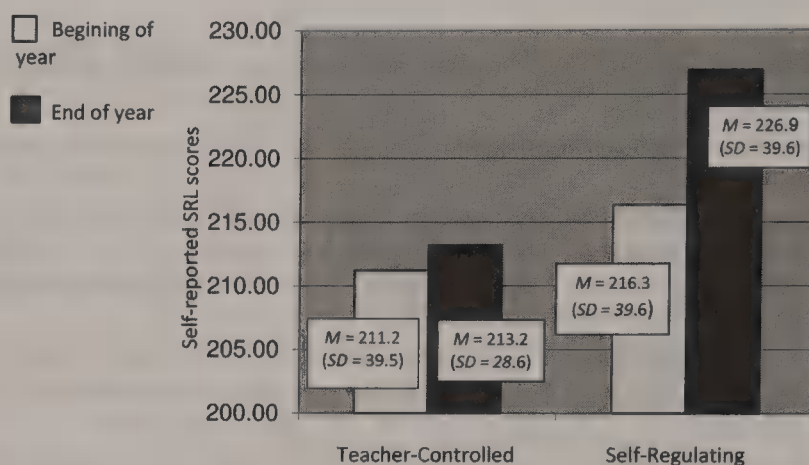


Figure 1. Pre- and posttest means (and *SDs*) for perceived self-reported SRL scores of TC and SR students (score range: 13–337 points).

Correlations Between Perceived Self-Reported (LASSI) and Enacted SRL and Science Knowledge. Pearson correlations were examined between perceived self-reported SRL and science knowledge in the SR and TC groups, at the beginning and the end of the year, as well as between enacted SRL and science knowledge in the SR group at the end of the year.

RESULTS

Perceived Self-Reported SRL in SR and TC Groups

As presented in Figure 1, self-reported SRL (LASSI) scores of SR students increased from the beginning of the intervention toward its end significantly more than those of TC students, $F(1,53) = 19.82, p < .001$, with a medium effect size, Cohen's $d = .40$. Namely, after experiencing active regulation during the academic year, SR students rated themselves significantly higher than their TC counterparts regarding the application of SRL.

Changes in Enacted SRL in the SR Group

Repeated-measures MANOVA showed a significant increase in students' enacted SRL scores over the entire academic year, $F(2,25) = 4.83, p < .001$. The largest increase in mean score was evidenced from the end of Term 1 ($M = 126.80, SD = 49.5$) to the end of Term 2 ($M = 140.70, SD = 48.5$). The mean score then remained stable through the end of Term 3 ($M = 138.60, SD = 48.2$). Post hoc tests of between-subjects effects showed a significant change from Term 1 to Term 2, $F(1,25) = 6.9, p < .01$, and a significant change from Term 1 to Term 3, $F(1,25) = 3.2, p < .05$, but no significant change from Term 2 to Term 3, $F(1,25) = .72, p > .05$.

The low scores of enacted SRL during Term 1 may be explained by students' process of familiarization with the three instruments and their habituation to the new SR environment. The significant increase in SRL enactment over the second term and its maintenance over the third term may reflect students' successful application of SR, probably due to their increasing mastery of the protocols and their growing experience in self-guided study. We next present examples of students' regulation activities taken from observation notes and student protocols.

Observed Management of SR Environment. Observations indicated that SR students repeatedly rearranged their learning environment. For example, in the beginning of the year, students' seating arrangement mirrored their traditional classroom setting, and students performed experiments at the location where the teacher placed materials despite cramped conditions. In time, students were observed choosing a preferred seat (e.g., near a friend, alone) and bringing other materials for learning to this selected location. From the second term on, observations indicated that an increasing number of students exhibited initiative in selecting specific locations relevant to different activity types. For instance, some selected a quiet corner for reading a textbook chapter, or moved to work with a friend for collaborative answering of textbook questions, whereas others answered them alone. Some were observed saying, "Be quiet; it is impossible to work in this noise," suggesting that students' increasingly adapted their learning environment to their preferred personal needs.

WSRI Analysis: Setting Weekly Goals. Students' goal phrasings changed very little over time in spite of their growing practice in goal setting and monitoring. Students in the SR group began the week by establishing weekly goals like "I have to work on Chapter 2 this week" (a performance goal). Such phrasing of goals was typically too general and vague to enable efficient planning and monitoring and was mostly based on the teacher's suggested weekly plan given in the YSRI. About 80% of students' weekly goals in all three terms were phrased as performance goals relating to curriculum coverage (e.g., "I have to finish the topic of Mendel"). Only 20% of goals were phrased as mastery goals, relating to knowing and understanding specific contents (e.g., "My weekly goal is to go over the materials again even though I already finished learning Chapter 4" or "I have to go back and learn the topics that I didn't understand in Chapter 5").

WSRI Analysis: Planning and Enactment. The protocol data revealed many instances of students' awareness of links between their personal learning preferences and their learning outcomes, as reflected in these students reasoning about the identified gaps and suggestions for closing them. Students' weekly planning involved selecting activities for attaining the stated goals, sequencing them, allocating time for the execution of each activity, and, if needed/desired, assigning homework. Students tended to select activities that were familiar to them from their traditional school learning experiences, such as reading the textbook chapters or answering its questions. Despite the SR group's long-term exposure to regulatory experiences, their selection of activity types changed very little over the year (see Table 6), with the exception of increases in help seeking from peers and the teacher. Analysis showed that once students discovered that talking with peers as well as seeking teachers' help was not only legitimate but also beneficial, their WSRI revealed increased planning and enactment of these activities, from the first to the third terms. Another salient point to be considered is that students encountered difficulties when completing the protocols and asked to define learning activities, and they were often satisfied with general phrasing ("learning"), as expressed in the high percentages of such activities presented in Table 6. In turn, this deficient ability hindered students' ability to improve their planning and monitoring.

Sequencing and time allocation were new experiences for students, and their enactment of plans revealed changes along the year. Enactment required students' adaptation of their initial plans to the dynamic environment of the classroom and to their own needs (e.g., time limitations, level of understanding, fatigue, boredom with a topic, low motivation), by changing some of the activities and readjusting time schedules. For example, one student planned to "read the page" for 15 minutes, then to "define concepts" for 10 minutes, and

TABLE 6
Number (and Percentage) of Planned and Enacted Activities Chosen by SR
Students in Each of the Three Terms (in WSRI Protocols)

Activity	Term 1		Term 2		Term 3	
	Planned	Enacted	Planned	Enacted	Planned	Enacted
Reading	54 (19%)	45 (19%)	21 (13%)	19 (13%)	12 (11%)	37 (27%)
Answering textbook questions	108 (37%)	63 (26%)	51 (32%)	41 (28%)	26 (24%)	30 (22%)
Summarizing, identifying main ideas, concept maps, etc.	81 (28%)	69 (29%)	27 (17%)	20 (14%)	15 (14%)	12 (9%)
Discussions with teacher or peers	0 (0%)	27 (11%)	6 (4%)	20 (14%)	7 (6%)	12 (9%)
General phrasing (e.g., learning)	45 (16%)	36 (15%)	54 (34%)	45 (31%)	50 (45%)	45 (33%)

then to “draw a concept map” for 10 minutes In his enactment, he “read the page” for 20 minutes, then “asked the teacher for explanations and we talked” for 5 minutes, and then he “wrote the main ideas” for 10 minutes Although the student planned to focus on new concepts while reading, his realization that he needed the teacher’s help resulted in changing his planned activities and managing time accordingly.

In their enactment column, most students reported staying on-task about 90% of the total lesson time, with only short off-task episodes reported (e.g., “chatting with friends on unrelated topics”). These data match the observation notes and were quite surprising. In this heterogeneous classroom, we expected that more students would take advantage of the provided freedom of the SR environment to avoid encounters with the difficulties presented by the biology and self-study. The importance of time as a valuable resource surfaced often in students’ reports, probably due to students’ routine reporting of time schedule, lesson time management, and monitoring of progress. Students identified explicit gaps easily. For example, they mentioned gaps between planning and enactment, whether regarding time schedules, sequences of activities, or activity types. However, frequently these gaps reflected various underlying difficulties that were not identified easily, like lacunae in their own available knowledge or a deficit in the abilities required to perform the activity. For example, a gap between the planned and enacted duration required to summarize a topic could result from deficient knowledge and understanding of this topic. Although students identified the gap in duration easily due to its visibility in the protocol, they seldom linked the gap to their own knowledge deficit. To close such gaps, students suggested modifications in behavior: “I changed my way of learning and now instead of answering the questions in the book I read with pauses and highlight the main ideas” or “I have found that it takes much less time” or “I know I have to take some work home.” Many students expressed awareness of the need to match their personal characteristics to their learning behaviors. For instance, they wrote, “I need to adjust my weekly time planning to my own pace and not to my peers’ pace” or “I must concentrate and ignore others’ talking because I can’t concentrate.” Students’ weekly examination of goals’ attainment as related to their performance (enactment of plans) strengthened their awareness of the links between actions and outcomes.

SR students determined their own learning pace. Managing time, they were aware of time loss and its implications for planning adjustments. For instance, in the first term one of the girls wrote, “I need to assign myself homework and hurry up.” In the following term, she wrote, “I am doing great, I am progressing according to the schedule and I understand.” In addition, their protocols reflected increased accountability for learning. Students wrote, “I wasn’t serious about my assignment” or “I need more sleep so I won’t be tired in class” or reinforced their performance when feeling satisfied (“WOW, I really get it now”), demonstrating essential SRL elements.

TSRI Analysis. In the SR group, students had to infer from the teacher’s written indication of incorrect responses on tests, without any explicit teacher discussion or explanation of errors. Some students did seem able to reach adequate conclusions about how to perform better in the future; for instance, when one student erred repeatedly in the *factual knowledge* part, she wrote, “I have to memorize the content.” Likewise, a student whose mistakes were mostly in the *understanding* part of the test wrote, “I should read deeper and insist on understanding what I read.” However, the latter student did not indicate the needed activities to promote reading depth, thereby decreasing his likelihood of improving his performance despite his awareness of the problem. Students’ vagueness of phrasing, like “I have to learn better” or “I have to pay more attention,” impeded their ability to monitor successfully.

Students’ Knowledge of Genetics in the SR and TC Groups

A MANOVA of all students’ science knowledge scores at the end of Term 1 (baseline) showed no significant difference between the SR group ($M = 59.84$) and the TC group ($M = 57.15$), $F(2,52) = 2.41$, $p > .05$, suggesting that these two heterogeneous groups of students began their study with similar knowledge of genetics, thus validating their comparison. In the beginning of Term 2 and again at the start of Term 3, all students’ prior knowledge of each new topic was tested. As expected, these pretests revealed very low science knowledge scores in both groups (see Table 7). The low mean scores at the posttests, too, in both groups (see Table 7) reflected these ninth graders’ difficulties in learning complex genetics topics, as reported previously by many researchers. However, a posttest MANOVA of the two groups’ scores at the end of Term 3 (the fifth interval) showed a significant group difference in students’ constructed genetic knowledge over the

TABLE 7
Means, Standard Deviations, *t* Tests, and Effect Sizes (Cohen’s *d*) for SR and TC Students’ Science Knowledge Scores Over the School Year

Test	Group				<i>t</i>	<i>d</i>
	SR (<i>n</i> = 25)		TC (<i>n</i> = 27)			
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		
Baseline (after Term 1)	59.84	7.87	57.15	10.91	.78	0.29
Pretest for Term 2	9.17	3.95	12.41	2.89	.84	0.94
Posttest for Term 2	63.42	10.78	47.81	12.25	1.34	1.35
Pretest for Term 3	28.17	7.22	20.18	4.98	1.44**	1.30
Posttest for Term 3	66.73	8.48	39.96	10.3	1.28**	2.85

Note: $d = X_f - X_m / 1/2(SD_f + SD_m)$. Score range: 0–100 points. ** $p < .01$.

Total score comprised 80% questions on understanding and application (qualitative learning) and 20% on declarative knowledge (memorizing information).

year, $F(2,52) = 2.24$, $p < .05$, in favor of the SR group. The increasing gap between the two student groups' knowledge of genetics was also reflected by the small effect size at the beginning of the year and the large effect size at its end.

We next present several examples of both student groups' science knowledge and understanding as expressed in their test responses. Examination of TC students' more frequent errors in multiple-choice knowledge questions, compared to SR students, revealed the TC students' less accurate body of knowledge on microlevel, abstract, symbolic, and complex concepts. For example, errors were related to concepts that included symbolic signs (e.g., describing the genotypes) or were composed of two adjacent words (e.g., homologous chromosomes). The advantage of the SR group regarding such complex knowledge may suggest that learning in the SR constituted an environment that promoted students' realization of their potential to learn.

An open-ended test example reveals the SC students' deeper scientific understanding than their TC peers despite similar difficulty in both groups in relating concurrently to different levels of the organism—the macrolevel as expressed in an individual's phenotype and the microlevel of the person's chromosomes. When asked to explain at what probability a healthy child would be born to two healthy parents if they already had a sick child with a chromosome X deficiency, many students could not respond at all to the question (59% of the TC group and 44% of the SR group), and only 3% of all those who responded could relate to probability. However, among those answering this question, a common reasoning in the SR group was "It is possible that these parents would have a healthy son if the mother is heterozygote and the relevant gene is a recessive one," whereas in the TC group students typically reasoned, "This situation is possible because if both parents are healthy the mother could pass the healthy gene to the healthy child."

Another example for differences in knowledge and understanding of the SR and TC students was revealed by the next question concerning the concept of homozygote: "Would an individual homozygote for a single trait be homozygote regarding all his other traits?" The percentage of students' correct responses was higher in the SR group than in the TC group (76% vs. 56%, respectively).

Correlation Between Perceived Self-Reported SRL and Science Knowledge. A similar baseline pattern of positive yet insignificant Pearson correlations emerged in the two groups between perceived self-reported SRL and genetics knowledge at the end of Term 1 (SR group: $r = .18$; TC group: $r < .22$; $p > .05$). In the posttest of the end of the last term, this positive yet insignificant correlation remained for the TC group ($r = .24$, $p > .05$), suggesting little correlation between academic achievement and perceived self-reported application of SRL. However, the SR students in the posttest showed a moderately significant correlation ($r = .46$, $p < .05$), exhibiting a significantly increased correlation over time as students' experiences with SRL grew.

Correlation Between Enacted SRL and Science Knowledge (SR Group Only). An insignificant baseline correlation emerged between enacted SRL and genetics knowledge at the end of Term 1 ($r = .37$, $p > .05$). A strong significant correlation between enacted SRL and science knowledge was found at the end of Term 2 ($r = .44$, $p < .05$). This correlation was stronger at the end of Term 3 ($r = .60$, $p < .01$). These findings regarding the strengthening correlation between actions and learning outcomes over time emphasizes the importance of a long-term intervention that allows most students to adjust to this new autonomous mode of learning. Enacted SRL was found to be a significant predictor of

science achievement scores, accounting for 26% of the variance ($R^2 = .26$, $F(1,25) = 7.355$, $p < .05$).

DISCUSSION AND FINAL THOUGHTS

The present long-term study investigated SR students' management of their learning processes in comparison to students who were taught in a TC environment. We focused on changes over time in ninth graders' SRL behaviors and their science knowledge. Findings showed that the SR students improved significantly over the school year as compared with their counterparts, both regarding science knowledge and perceived self-reported SRL. The SR group significantly improved their enacted SRL over the year. These findings suggest that autonomous learning while repeatedly enacting SRL in the dynamic classroom context is effective in promoting meaningful science learning.

Changes in Perceived Self-Reported and Enacted SRL Over the Year

Corroborating many prior research reports (e.g., Azevedo & Cromley, 2004; Eilam & Aharon, 2003; Schraw et al., 2006), the present study showed that students who were provided with the opportunity to regulate their own learning improved both their perceived self-reported SRL and their enacted SRL over time and in turn, their genetics knowledge. Differently from the SR students, no significant changes were found in perceived self-reported SRL scores among the TC students, probably because of the external regulation exerted by the teacher. Owing to the reported relations between students' perceived self-reported SRL and their enacted SRL (Eilam & Aharon, 2003), we may assume that if perceived self-reported SRL does not change over time, enactment of regulation does not change either.

Similarly, the SR group of students constructed more knowledge and understanding and showed better application of that knowledge than their TC counterparts. These results are not surprising in light of many researchers' reports that active learning, where students are deeply engaged in the process of knowledge construction, promotes learning (Mayer, 2008), thereby meeting the growing demands of science education goals (Schraw et al., 2006).

The SR group's improvement in enacting SRL was found to be related positively to students' academic science knowledge, as reported in other studies (Lee et al., 2009; Perels et al., 2009; Winne, 2001). These results may be explained by the metacognitive aspects of SRL. Through the long-term self-study, during which students improved their SRL application, they became aware of a certain extent of their own knowledge, abilities, and preferred environments. They learned to control and manage different aspects of their learning (as expressed in their increased reported and enacted SRL). Improving SRL means that students who fail to improve achievement may make changes in their learning processes until they reach higher achievement. Facing all aspects of long-term autonomous learning for the first time, the SR students made a transition from the TC environment that was familiar from past learning to the current SR environment, which necessitated a shift in their learning conceptions. Students had to exhibit full autonomy in making decisions concerning learning goals, information gathering and processing, application of strategies, monitoring processes, and pace, while accepting full responsibility for their actions. This shift in conceptions of learning required time as well as repeated but diverse experiences. Given repeated opportunities for self-determined planning, performance, and self-evaluation—students improved their general regulation levels over time as exhibited by their enacted

SRL scores. They also strengthened the relations between their learning processes and products expressed in increased correlations between SRL and achievement over the year.

In contrast, no significant change was found in the perceived self-reported SRL of the TC students, whose activities and learning environment were regulated by the teacher and did not necessarily match their preferences, knowledge, and abilities. They were provided with limited opportunities to regulate their learning; hence, the TC group could not significantly change their posttest ranking of statements on the LASSI scales.

Our study is novel due to the unique learning environment. This complex environment encompasses different, interacting components that influence its function on many levels. In order for learning in such environment to be productive, these components should work in concert to fit individuals' needs and preferences. Three of the components contributed to this environment's uniqueness in particular:

1. Diverse opportunities to practice SR were repeated over the year, due to the specific task characteristics, namely, the need to learn genetics contents, which involved the ability to construct understanding of new concepts, to apply these concepts in different contexts, and to solve related problems. These abilities were acquired, applied, and refined repeatedly along the year. The gained knowledge was applied in relation to diverse situations involving the macro- and microlevels of phenomena, concrete and abstract concepts and processes, as well as statistics. Such learning experiences evolved from many cycles of setting goals, planning activities to achieve them, monitoring advancement toward the distant goals of constructing knowledge and understanding of genetics, and refining learning modes, strategies used, and time management in light of previous experiences. These cycles were developed applying Zimmerman's (2000) model of SRL as a process, including its three phases—forethought, performance, and self-reflection.
2. The learning environment aimed not only to enable students to practice SRL but also to promote their related abilities, and in particular planning and monitoring. It promoted students' understanding of the links between strategy types and outcomes, alongside the recognition that these links may change in different contexts/environments. Such a realization makes students examine the conditions under which they act in a certain way rather than acting spontaneously. They become more aware of cues, and they increase their ability to perceive cues from the environment, from themselves, or from the social context. Such awareness, in turn, increases metacognitive thinking and may result in a refined mental model of the SRL process and the knowledge constructed. This particular way of promoting students' SRL abilities was initiated by the instruments developed for use in this environment. These instruments supported students' long-term self-study of complex contents, constituted a platform that on the one hand increased their awareness of the possible influences of various factors on the outcomes and on the other hand explicated different relations between all involved factors. Such practices may promote mental flexibility that allows for adaptation to new environmental demands.

Analysis of students' protocols exhibited changes in their behaviors over time as expressed by the scoring of the protocols and as supported by the observations. For example, some students worked alone for three lessons and then worked together in the following two lessons, discussing their insights, and comparing their notes. Such changes in behavior which fits the content learn suggest self-regulation. This analysis also revealed some salient deficiencies in students' ability to plan. For example, students needed to mindfully select a goal while considering the specific learning context and the need to sharpen the goal

definition in a way that would elicit a list of activities (a plan) for achieving this goal that would not hinder monitoring. However, students exhibited mostly rote selection of goals according to textbook chapters, and they phrased these goals as performance goals rather than mastery ones (Locke & Latham, 2006). This finding suggests that students were still bound by the textbook or the teacher's suggestions, with limited utilization of the potential inherent in the autonomy granted to them. In addition, the vague phrasing of the goals evidenced in the protocols limited the repertoire of activities suggested by students for achieving this goal of knowing and understanding. Most of the activities they enacted in their process of learning were those practiced routinely at school, in spite of the autonomy granted. It seems that school learning modes are deeply rooted in students' behaviors, in ways that impede their ability to fully perceive new learning potential in the situations they encounter. We suggest here that the development of tools and the provision of opportunities are not enough for developing SR students. It is also necessary to increase these students' awareness of the many new and ever developing possibilities inherent in their encountered surroundings, thus developing their abilities to cope successfully.

3. The longitudinal nature of this study enabled us to trace changes in students' SRL and learning outcomes. The increase found in Term 2 for SR students' scores in enacted SRL over time, which was maintained over Term 3, strengthened our claim for the importance of a long-term intervention. Students' process of adaptation to the novel learning environment, following many years of teacher control, was time consuming. The fruits of students' mastery of the given tools and efforts to adapt to the SRL environment became noticeable only at the end of the second term, during which students habituated an individual learning process, and they continued applying SRL throughout the third term. A similar delayed effect was reported by other studies (e.g., Mevarech & Amrany, 2008).

It should be emphasized in this ecologically valid study that the examination and measurement of students' perceived self-reported and enacted SRL were carried out in heterogeneous classrooms with regard to students' learning of the regular national genetics curriculum. Although many previous studies examined the effect of SRL on students' learning of school materials, these were mostly short-term studies of limited tasks. Knowledge regarding students' self-study of a whole-year curriculum, which is most important for promoting learning in the public schools, is still deficient. This is another special aspect of the present study.

Knowledge of Genetics

Although the two student groups' tests exhibited similar knowledge of genetics at the outset of the intervention, the SR students revealed significant improvements in science knowledge over the year, with a growing effect size, compared to the TC students. This result is striking because the complexity of the studied science contents increased from term to term, with concepts and principles becoming more abstract and microlevel; yet, greater complexity correlated with larger gaps between the two groups' science knowledge levels, suggesting that students' learning mode was a core factor influencing achievement.

However, as mentioned above, improvement in science knowledge construction may evolve not only from improvement in SRL but also from students' active learning mode and deep involvement in their learning processes (Mayer, 2008). There is no question that students involved in autonomous learning are generally more mentally active than in a TC classroom. However, in the present case we do not have data to confirm this.

In sum, teachers often complain about deficient time, students' inattention, or other difficulties characterizing heterogeneous classrooms. One beneficial solution is to transfer responsibility of learning to students, making them accountable for its products (Eilam & Aharon, 2003; Van den Hurk, 2006). Inasmuch as SRL is correlated with achievement, all students should be provided with opportunities to practice it systematically in diverse subject matters and situations, and to explicitly acquire its many skills. Such experiences may improve students' epistemological beliefs about learning and may promote their ability to become lifelong learners (Eshel & Kohavi, 2003; Puustinen & Pukkinen, 2001; Rozendaal, Minnaert, & Boekaerts, 2005). However, successful implementation of SRL in schools requires the training of teachers. This study highlighted the need for more long-term research on SRL in authentic classrooms and for the design of diverse models for implementing SRL pedagogy in students' everyday learning environments.

Study Limitations

We focused primarily on the overall picture and did not delve into the qualitative descriptions yielded by students' protocols and by their responses to the open-ended science knowledge questions. Moreover, future researchers would do well to complement the current findings with observations, traces, video data, and interviews concerning students' enacted SRL along the year, in addition to their written protocols, to deepen understanding of students' many variations of learning regulation and give insights into their thinking. Despite their importance, long-term studies of SRL held in the ecologically valid setting of authentic classroom learning conditions are few (Paris & Paris, 2001; Puustinen & Pukkinen, 2001). Such studies increase the number of possible intervening factors, increase the difficulties involved in data measurement and collection, and limit our ability to deeply examine each student and each of the SRL components involved. Such studies are greatly in need to extend knowledge of students' SRL beyond the lab and thereby to inform practice. The importance of our findings lies in the current recognition that learning, knowledge, and context are inseparable (Barab & Squire, 2004). Such limitations were expressed, for example, in our inability to differentiate the effect of SRL from that of active learning or to differentiate the SRL effects in the two learning modes.

The authors thank Dee B. Ankonina for her editing contribution.

REFERENCES

- Abrahams, I., & Reiss, M. J. (2012). Practical work: Its effectiveness in primary and secondary schools in England. *Journal of Research in Science Teaching*, 49(8), 1035–1055.
- Aharon, I. (2003). Personality traits as predictors of self-regulated learning in junior high school students. Unpublished master's thesis, University of Haifa, Israel (in Hebrew).
- Allen, J. F., Hendler, J., & Tate, A. (1990). *Readings in planning*. San Francisco: Kaufman.
- Apple, M. W. (2008). Curriculum planning. Content, form, and the politics of accountability. In F. M. Connelly (Ed.), *The Sage handbook of curriculum and instruction* (Chap. 2, pp. 25–44). Thousand Oaks, CA: Sage.
- Azevedo, R. (2007). Understanding the complex nature of self-regulatory processes in learning with computer-based learning environments: An introduction. *Metacognition Learning*, 2, 57–65.
- Azevedo, R. (2009). Theoretical, conceptual, methodological, and instructional issues in research on metacognition and self-regulated learning: A discussion. *Metacognition Learning*, 4, 87–95.
- Azevedo, R., & Cromley, J. G. (2004). Does training on self-regulated learning facilitate students' learning with hypermedia? *Journal of Educational Psychology*, 96(3), 523–535.
- Bannert, M., & Mengelkamp, C. (2007). Assessment of metacognitive skills by means of instruction to think aloud and reflect when prompted. Does verbalization method affect learning? *Metacognition Learning*, 3, 39–58.

- Barab, S. A., & Squire, K. (2004). Design-based research: Putting a stake in the ground. *Journal of the Learning Sciences*, 13(1), 1–14.
- Bell, B. S., & Kozlowski, S. W. J. (2002). Goal orientation and ability: Interactive effect on self-efficacy, performance, and knowledge. Retrieved December 18, 2013, from Cornell University School of Industrial and Labor Web site: <http://digitalcommons.ilr.cornell.edu/hrpubs/9/>.
- Bell, P., Davis, E. A., & Linn, M. C. (1995). The knowledge integration environment: Theory and design. *Proceedings of the Computer Supported Collaborative Learning Conference (CSCL '95)*, Bloomington, IN (pp. 14–21). Mahwah, NJ: Erlbaum.
- Berthold, K., Nuckles, M., & Renkl, A. (2007). Do learning protocols support learning strategies and outcomes? The role of cognitive and metacognitive prompts. *Learning and Instruction*, 17, 564–577.
- Bråten, I., & Samuelstuen, M. S. (2007). Measuring strategic processing: Comparing task-specific self-reports to traces. *Metacognition Learning*, 2(1), 1–20.
- Bronsan, T. (1990). Categorizing macro and micro explanation of material change. In P. L. Lijnse, P. Licht, W. de Vos, & A. J. Vaarlo (Eds.), *Relating macroscopic phenomena to microscopic particles* (pp. 198–211). Utrecht, The Netherlands: CD Press.
- Brown, R., & Pressley, M. (1994). Self-regulated reading and getting meaning from text: The transactional strategies instructional model and its ongoing validation. In D. Schunk & B. Zimmerman (Eds.), *Self-regulation of learning and performance: Issues and educational applications* (pp. 155–180). Hillsdale, NJ: Erlbaum.
- Butler, D. L., & Wine, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, 65(3), 245–281.
- Cano, F. (2006). An in-depth analysis of the Learning and Study Strategies Inventory (LASSI). *Educational and Psychology Measurement*, 66(6), 1023–1038.
- Carver, C. S., & Scheier, M. F. (1990). Origins and functions of positive and negative affect: A control-process view. *Psychological Review*, 97, 19–35.
- Chi, M. T. H., DeLeeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, 18, 439–477.
- Cleary, T. J., & Labuhn, A. S. (2013). Application of cyclical self-regulation intervention in science-based context. In H. Bembunty, T. J. Cleary, & A. Kitsantas (Eds.), *Application of self-regulated learning across diverse disciplines: A tribute to Barry Zimmerman* (Chap. 4, pp. 89–124). Charlotte, NC: Information Age.
- Davis, E. A., & Linn, M. C. (2000). Scaffolding students' knowledge integration: Prompts for reflection in KIE. *International Journal of Science Education*, 22(8), 819–837.
- DeLeeuw, N., & Chi, M. (2003). Self-explanation: Enriching a situation model or repairing a domain model? In G. Sinatra & P. Pintrich (Eds.), *Intentional conceptional change* (pp. 55–78). Mahwah, NJ: Erlbaum.
- Dignath, C., Buettner, G., & Langfeldt, H. P. (2008). How can primary school students learn self-regulated learning strategies most effectively? A meta-analysis of self-regulation training programs. *Educational Research Review*, 3(2), 101–129.
- Duncheon, J. C., & Tierney, W. G. (2013). Changing conceptions of time: Implications for educational research and practice. *Review of Educational Research*, 83(2), 236–272.
- Eilam, B. (2002). Strata of comprehending ecology: Looking through the prism of feeding relations. *Science Education*, 86, 645–671.
- Eilam, B. (2012a). System thinking and feeding relations: Learning with a live ecosystem model. *Instructional Science*, 40(2), 213–239.
- Eilam, B. (2012b). Possible constraints of visualization in Biology: Challenges in learning with multi-media. In D. Treagust & C.-Y. Tsui (Eds.), *Multiple representations in Biological education* (Chap. 4, pp. 55–73). Models and modeling in science education. Berlin: Springer.
- Eilam, B. (2012c). Teaching, learning, and visual literacy: The dual role of visual representation in the teaching profession. New York: Cambridge University Press.
- Eilam, B., & Aharon, I. (2003). Students' planning in the process of self-regulated learning. *Contemporary Educational Psychology*, 28, 304–334.
- Eilam, B., Zeidner, M., & Aharon, I. (2009). Student conscientiousness, self-regulated learning, and science achievement: A prospective field study. *Psychology in the Schools*, 46(5), 420–432.
- Elliot, A. J. (1999). Approach and avoidance motivation and achievement goals. *Educational Psychologist*, 34(3), 169–189.
- Eshel, Y., & Kohavi, R. (2003). Perceived classroom control, self-regulated learning strategies and academic achievement. *Educational Psychology*, 23(3), 249–260.
- Ge, X., Chen, C. H., & Davis, K. A. (2005). Scaffolding novice instructional designers' problem-solving processes using question prompts in a web-based learning environment. *Journal of Educational Computing Research*, 33(2), 219–248.

- Gollwitzer, P. M. (1996). The volitional benefits of planning. In P. M. Gollwitzer & J. Bargh (Eds.), *The psychology of action: Linking cognition and motivation to behavior* (pp. 287–312). New York: Guilford Press.
- Greene, B. A., & Land, S. M. (2000). A qualitative analysis of scaffolding use in a resource-based learning environment involving with the world wide web. *Journal of Educational Computing Research*, 23(2), 151–180.
- Greene, J. A., & Azevedo, R. (2007). A theoretical review of Winne and Hadwin's model of self-regulated learning: New perspectives and directions. *Review of Educational Research*, 77(3), 334–372.
- Hacker, D. J. (1998). Definitions and empirical foundations. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 1–23). Mahwah, NJ: Erlbaum.
- Jordan, R. C., Ruibal-Villasenor, M., Hmelo-Silver, C. E., & Etkina, E. (2011). Laboratory materials: Affordances or constraints? *Journal of Research in Science Teaching*, 48(9), 1010–1025.
- Kapa, E. (2007). Transfer from structured to open-ended problem solving in a computerized metacognitive environment. *Learning and Instruction*, 17, 688–707.
- Kiewra, K. A. (2002). How classroom teachers can help students learn and teach them how to learn. *Theory into Practice*, 41(2), 71–80.
- Lee, H. W., Lim, K. Y., & Grabowski, B. (2009). Generative learning strategies and metacognitive feedback to facilitate comprehension of complex science topics and self-regulation. *Journal of Educational Multimedia and Hypermedia*, 18(1), 5–25.
- Lemke, J. L. (2002). Science and experience. In J. Wallace & W. Louden (Eds.), *Dilemmas of science teaching: Perspectives on problems of practice* (Chap. 1, pp. 30–33). London: Routledge/Falmer.
- Lewis, J., & Kattman, U. (2004). Traits, genes, particles and information: Re-visiting students' understandings of genetics. *International Journal of Science Education*, 26(2), 195–206.
- Lin, L.-M., & Zabrocky, K. (1998). Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology*, 23, 345–391.
- Lin, X. (2001). Designing metacognitive activities. *Educational Technology Research and Development*, 49(2), 23–40.
- Linn, M. C., & Eylon, B.-S. (2006). Science education: Integrating views of learning and instruction. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 511–544). Mahwah, NJ: Erlbaum.
- Locke, E. A., & Latham, G. P. (2006). New directions in goal-setting theory. *Current Directions in Psychological Science*, 15(5), 265–268.
- Manlove, S., Lazonder, A. W., & de Jong, T. (2007). Software scaffolds to promote regulation during scientific inquiry learning. *Metacognition & Learning*, 2, 141–155.
- Marbach-Ad, G., & Stavy, R. (2000). Students' cellular and molecular explanations of genetic phenomena. *Journal of Biological Education*, 34(4), 200–205.
- Mayer, R. E. (2008). *Learning and instruction* (2nd ed.). Upper Saddle River, NJ: Pearson.
- Mevarech, Z. R., & Amrany, C. (2008). Immediate and delayed effects of meta-cognitive instruction on regulation of cognition and mathematics achievement. *Metacognition and Learning*, 33(2), 147–157.
- Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect." *Psychological Science*, 2(4), 267–270.
- Nicol, D. J., & Mcfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199–218.
- Nückles, M., Hübner, S., & Renkl, A. (2009). Enhancing self-regulated learning by writing learning protocols. *Learning and Instruction*, 19(3), 259–271.
- Papaleontiou-Louca, E. (2003). The concept and instruction of metacognition. *Teacher Development*, 7(1), 9–30.
- Paris, S. G., & Paris, A. H. (2001). Classroom applications of research on self-regulated learning. *Educational Psychologist*, 36(2), 89–101.
- Perels, F., Dignath, C., & Schmitz, B. (2009). Is it possible to improve mathematical achievement by means of self-regulation strategies? Evaluation of an intervention in regular math classes. *European Journal of Psychology of Education*, 24(1), 17–31.
- Perkins, D. (1998). What is understanding? In M. S. Wiske (Ed.), *Teaching for understanding: Linking research with practice* (pp. 39–57). San Francisco: Jossey-Bass.
- Peters, E., & Kitsantas, A. (2010). The effect of nature of science metacognitive prompts on science students' content and nature of science knowledge, metacognition, and self-regulatory efficacy. *School Science and Mathematics*, 110(8), 382–396.
- Pieschl, S. (2009). Metacognitive calibration: An extended conceptualization and potential applications. *Metacognition Learning*, 6, 205–211.
- Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In Boekaerts, M., Pintrich, P., & Zeidner, M. (Eds.), *Handbook of self-regulation* (Chap. 14, pp. 451–502). San Diego, CA: Academic Press.

- Pintrich, P. R., & Schunk, D. H. (1996). *Motivation in education: Theory, research, and applications*. Englewood Cliffs, NJ: Merrill-Prentice-Hall.
- Prins, F. J. (2002). Search & see: The roles of metacognitive skillfulness and intellectual ability during novice inductive learning in a complex computer-simulated environment. Leiden, The Netherlands: Print Partners Ipskamp.
- Puustinen, M., & Pulkkinen, L. (2001). Models of self-regulated learning: A review. *Scandinavian Journal of Educational Research*, 45(3), 269–286.
- Roth, W-M., & Lee, S. (2004). Science education as/for participation in the community. *Science Education*, 88, 263–291.
- Rozendaal, J. S., Minnaert, A., & Boekaerts, M. (2005). The influence of teacher perceived administration of self-regulated learning on students' motivation and information processing. *Learning and Instruction*, 15, 141–160.
- Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1), 68–78.
- Schraw, G., Crippen, K. J., & Hartley, K. (2006). Promoting self-regulation in science education: Metacognition as part of a broader perspective on learning. *Research in Science Education*, 36(1), 111–139.
- Schunk, D. H. (1990). Goal setting and self-efficacy during self-regulated learning. *Educational Psychologist*, 25(1), 71–86.
- Spiro, R. J., Feltovich, P. J., Jacobson, M. J., & Coulson, R. L. (1991). Cognitive flexibility, constructivism, and hypertext: Random access instruction for advanced knowledge acquisition in ill-structured domains. *Educational Technology*, 31(5), 24–33.
- Stinner, A. (1992). Science textbooks and science teaching: From logic to evidence. *Science Education*, 76, 1–16.
- Van den Hurk, M. M. (2006). The relation between self-regulated strategies and individual study time, prepared participation and achievement in a problem-based curriculum. *Active Learning in Higher Education*, 7, 155–169.
- Van der Veen, I., & Peetsma, T. (2009). The development in self-regulated learning behaviour of first-year students in the lowest level of secondary school in the Netherlands. *Learning and Individual Differences*, 19(1), 34–46.
- van Velzen, J. H. (2013). Students' explanations of their knowledge of learning processes. *Educational Studies*, 39(1), 83–95.
- Veenman, M. V. J. (2007). The assessment and instruction of self-regulation in computer-based environments: A discussion. *Metacognition Learning*, 2, 177–183.
- Veenman, M. V. J. (2011). Alternative assessment of strategy use with self-report instrument: A discussion. *Metacognitive Learning*, 6, 205–2011.
- Weinstein, C. E., Palmer, D. R., & Schulte, A. C. (1987). *Learning and Study Strategies Inventory*. Clearwater, FL: H & H.
- Winne, P. H. (1995). Self-regulation is ubiquitous but its forms vary with knowledge. *Educational Psychologist*, 30, 223–228.
- Winne, P. H. (2001). Self-regulated learning viewed from models of information processing. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement: Theoretical perspectives* (pp. 153–189). Mahwah, NJ: Erlbaum.
- Zimmerman, B. J. (2000). Attaining self-regulation. A social-cognitive perspective. In Boekaerts, M., Pintrich, P., & Zeidner, M. (Eds.), *Handbook of self-regulation* (Chap. 2, pp. 13–39). San Diego, CA: Academic Press.
- Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory into Practice*, 41(2), 64–70.
- Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Research Journal*, 45(1), 166–183.
- Zimmerman, B. J., Greenberg, D., & Weinstein, C. E. (1994). Self-regulating academic study time: A strategic approach. In D. H. Schunk & B. J. Zimmerman (Eds.), *Self-regulating of learning and performance. Issues and educational applications* (pp. 181–199). Hillsdale, NJ: Erlbaum.

Trying Biology: The Scopes Trial, Textbooks, and the Antievolution Movement in American Schools, by Adam R. Shapiro. University of Chicago Press: Chicago, IL, USA, 2013. 193 pp. ISBN 978-0-226-02945-0.

The State of Tennessee vs. John Thomas Scopes, commonly referred to as the “Scopes trial,” was a pivotal event in the history of evolution education in the United States. It was the first legal proceeding about teaching evolution in public schools to gain national attention. The Butler bill, a newly approved Tennessee law, banned any instruction denying the divine creation of humans or suggesting that they were descended from “lower” animals. In response, the American Civil Liberties Union announced their legal support for any educator willing to teach evolution in the public schools. John Scopes, a substitute high-school biology teacher, did just that. What ensued was a public spectacle involving the passionate arguments of William Jennings Bryan, a well-known politician and passionate antievolutionist, and Clarence Darrow, a prominent attorney who defended Scopes in Dayton, Tennessee. Though the trial is highly referenced in discussions of evolution education and has gained significant recognition over the years, Adam Shapiro’s *Trying Biology: The Scopes Trial, Textbooks, and the Antievolution Movement in American Schools* (2013) lends a fresh perspective by examining the social and political forces surrounding this historical event.

Shapiro integrates discussions of several recurring factors that shaped both the outcome and perceptions of the Scopes trial throughout the book. For example, he examines the influences of education bills being considered in the state legislature and practices of textbook companies during that time. Governor Austin Peay’s signing of the Butler bill was, in part, a negotiating tactic to pass a more general education bill. This bill aimed to make education compulsory and establish an eight-month school year. The notion of compulsory education was not especially popular among rural residents, who saw it as an attempt to change their culture and way of life. In tandem with this effort, textbook companies formed a monopoly in certain regions, leading to inflated book prices. In response, Tennessee passed a uniform textbook law and entered into five-year contracts with textbook publishers to make books more affordable. This set the stage for the state’s textbook commission to choose a biology textbook. Their choice was *Civic Biology*, written by George William Hunter and used by Scopes in his teaching. As Shapiro explains, although evolution content was included in the book, what made the curriculum of *Civic Biology* problematic was an overall focus on civic responsibility and economic relationships that were perceived to reflect urban environments more than rural ones:

Civic biology taught students to prepare for a life away from their traditional upbringing. Consequently, parents took exception to the presence of biology as well as to its content.

The fact that the books taught the historical development of species was a small concern. The overall discipline of civic biology and the presence of new schools intended to bring social progress were much more objectionable. (p. 83)

What we learn from Shapiro's discussion is that the Butler bill may have been more of a reaction to concerns about protecting local culture than it was about advancing a strong antievolutionist stance.

Along with the general education bill and corrupt practices of textbook companies, Shapiro explores how the Scopes trial was framed. At first, the defense planned to make the argument that teaching evolution was not equivalent to denying the teachings of the Bible; that it was possible to accept evolution within an overarching theistic framework. However, the true goal of the defense was to challenge the Butler bill in a higher court. With the assumption of a guilty verdict, it became easy to engage in a more direct conversation about the truth of evolution and religion. Darrow's famous cross-examination of Bryan, in which he questioned Bryan on the truth of the Bible, was widely reported in the media and shifted the narrative from a simple case of a law violation to a public debate pitting evolution and religion directly against one another.

Shapiro also discusses the response of textbook companies to the Scopes trial. Despite the author's clear objection, the American Book Company decided to de-emphasize evolution in the next edition of *Civic Biology*. One theme that is revisited throughout the book is the notion that the focus on selling textbooks trumped concerns over the content. Book companies had to devise strategies for removing evolution content in a way that would sell in the south without drastically changing the text to conserve reprinting costs. Key strategies included removing the word, "evolution," while still retaining some related ideas such as heredity and variation, and de-emphasizing the notion of humans being products of a common descent. It is interesting to note that avoiding the use of the "e-word" and excluding human evolution are strategies used nearly 90 years later by biology teachers today.

While I found *Trying Biology* to be an interesting and helpful read, it is likely better suited for individuals looking to supplement their knowledge on the historical forces that have shaped evolution education rather than for educators seeking practical implications for evolution instruction or educational reform. Though the claims are well substantiated, readers may find the discussion regarding the development of the textbook industry in the first half of the book somewhat tedious. In addition, I believe that more explicit consideration could have been devoted to how this reexamination of the Scopes trial relates to the status of evolution education today, and ways we might move forward.

Shapiro's account has certainly caused me to rethink this infamous trial and its influence on evolution education. The Scopes trial has been envisioned as a reflection of an America that was deeply divided over the issue of evolution. In contrast, Shapiro's research suggests that concerns had more to do with increased state control in the handling of educational affairs, particularly with the advent of compulsory education. With any trial, the narrative is often focused on who won and who lost. One might consider the Scopes trial as a loss for evolution because Scopes was found guilty and evolution was subsequently de-emphasized in textbooks, or alternatively as a win because it set the stage for a national conversation about the importance of including evolution in the biology curriculum. A discussion of winning and losing, however, is most likely a net loss for evolution education because it precludes thoughtful discussions with stakeholders about why it is a useful framework for understanding the natural world and how concerns about teaching it in the classroom could be alleviated in a respectful way. Gaining a deeper understanding of the key historical events of evolution education, such as the Scopes trial, helps us recognize the social and

political forces involved with educational decisions. Shapiro's examination of textbook companies, compulsory education, concerns of losing a sense of local values, personal agendas of attorneys, and oversimplifications reported in the media reveals a trial with all of the complexity typical of human interactions and very little support for a strict "science vs. religion" debate that persists in the public discourse. When advocating the teaching of evolution, stakeholders need to be aware of the social, political, and emotional nuances involved. Shapiro's *Trying Biology* helps us understand those nuances.

AARON J. SICKEL

*Department of Teacher
Education*

Ohio University

Athens, OH 45701, USA

DOI 10.1002/sce.21102

Published online 30 January 2014 in Wiley Online Library (wileyonlinelibrary.com).

I Died for Beauty: Dorothy Wrinch and the Culture of Science, by Marjorie Senechal. Oxford University Press, New York, NY, USA, 2013. ix + 300 pp. ISBN 978-0-19-973259-3.

In *I Died for Beauty: Dorothy Wrinch and the Cultures of Science*, Professor Emerita of Mathematics and History of Science and Technology Marjorie Senechal introduces us to a sometimes difficult, sometimes controversial, always brilliant mathematician and biochemist, Dr. Dorothy Wrinch. Senechal provides a historical account of Wrinch's life, bringing together stories derived from Senechal's own interactions with the scholar, interviews, written personal documents, published biographies, and archives in the United States, Canada, and England. The book provides science educators with an account of one of the first women to forge a career in a male-dominated profession: an account that provides insights into the history of mathematics and science, as well as women's changing roles in those fields.

Senechal begins by introducing us to a 76-year-old Wrinch. A senior professor living alone in a small apartment at Smith College in the year 1970, still entangled in her lifelong pursuit of the inner logic of protein molecules. Senechal worked as Dorothy's unpaid assistant making models and drawing illustrations for a book. This was the starting point of a relationship that would last until Wrinch's death in 1976. Senechal organized Wrinch's papers for an archivist after her death. Those personal and professional papers, and ultimately many follow up sources of information, are used to launch us on a journey into Dorothy Wrinch's past and catch glimpses of history in the making.

Dorothy Wrinch was born to English citizens residing in Argentina in 1894. Soon after her birth, the family returned to England, where Dorothy was educated. Born in a time of calls for quality education for women, a young Wrinch encountered people destined to make an impact on the world. The education provided by these trail blazers for women's rights, as well as Dorothy's own unyielding determination, led to her becoming instrumental in discovering the inner workings of protein molecules. Through Wrinch's story and those of the many people in her life, we witness history in science and math. She was a student and lifelong friend of Bertrand Russell, Nobel Prize winner in Literature, and a philosopher widely known for his work on mathematical logic. She engaged in a public debate with Linus Pauling, Nobel Prize winner in Chemistry. She was a friend and colleague of Dr.

D'Arcy Thompson, author of *On Growth and Form*, and Dorothy Hodgkin, Nobel Prize winner in Chemistry.

The quote "Well-behaved women seldom make history . . ." (Ulrich, 1976, p. 20) came to my mind often as I read the book. Dorothy Wrinch was not considered a well-behaved woman in her time and she did make history. Senechal introduces us to a stubborn woman focused on the abstract geometrical forms of proteins. Wrinch made major contributions to science and mathematics. Among these contributions, she was the first woman to earn a doctor of science degree from Oxford University; she was the first female mathematician to lecture to male students at Cambridge; she was a major contributor to our understandings of protein structure, and in the process, she pointed crystallographers in a new direction. Wrinch could be pushy. Once, when Girton offered her a mathematical apprenticeship with a modest stipend, instead of being grateful that they offered support for a woman to study mathematics, she drew up a table of her anticipated expenses and demonstrated that she needed a larger stipend. Wrinch was often stubborn. Even when empirical evidence was stacking up against her cyclol hypothesis, she continued to promote it. Wrinch did not shy away from controversy. She publicly feuded with Linus Pauling. She was also involved in a scandal at Amherst when her second husband, Glaser, was forced into early retirement for improperly using some of his grant funds to support her work.

Important to the book is the fact that its author, Senechal, is herself a scholar in this field, a woman in a still male-dominated profession and only one generation removed from Wrinch and her impacts on the profession. Senechal writes with an understanding of the professional practice, the content, and the experience of being a woman in math and science. Her admiration for Wrinch, and her enthusiasm for science and the beauty found within, makes this more than a traditional biography of a woman in science and mathematics.

Stylistically, Senechal's expressed desire is to provide us with a kaleidoscope of a book. I discovered that, at times, it can be frustrating to read a kaleidoscope of a book. At certain points, the book focuses on topics I expected: Dorothy's education, her life, and her work. Yet, at other points, I found myself reading about the lives of one or more of Wrinch's colleagues, or some mathematical concept, or scientific concept, history of science, poetry, and even some fiction. I also found myself in the basement of Smith College discovering lantern slides and learning about Senechal's career. The overall story is not conveyed in a linear manner and often, a change in focus is abrupt and unexpected. The narrative weaves in and out of time periods, lives, and subjects. Yet, in the end, the pieces came together for me and I knew a great deal more about this brilliant woman in the context of the history of mathematics and science. I was reminded that history isn't disconnected from the present and a person's life is intimately connected to the culture. I witnessed this through the stories about people, society, scientific concepts, as well as those lantern slides still lingering in the basement.

One thing I would have liked to find in *I Died for Beauty: Dorothy Wrinch and the Cultures of Science* is a deeper, more explicit discussion on the cultures of science. I believe that this is one of the configurations that was intended to emerge in the kaleidoscope approach; yet, I did not find it to be substantial. The relationship between Wrinch and the culture of science is always there, but only as a backdrop to the many intellectual, creative, and often colorful individuals, as well as science and mathematics. I believe that this is a missed opportunity to further the discussion on the challenges female scientists faced in their positions outside the *culture of power*. The culture of power elevates a group of people to a position where they have more control of money, people, and societal values than their non-culture-of-power peers (Delpit, 1988). Discussions of the culture of power in science are important to science education; yet authentic examples are few. Throughout this book, we see examples of how Wrinch's studies were influenced by the fact that, as a woman, she was outside the culture

of power in science and mathematics. What she studied was influenced by the fact that she could attend lectures at Cambridge only if the male professors allowed it. Her mentors had to be men that agreed to work with women; thus, she was destined to study with a group of males with certain beliefs about women and science. In addition, her approaches were often devalued for they differed from the norm. She spoke of fabrics of proteins, cyclol structures, and her work was nonempirical. She often worked without the needed assistance and had to support herself and her work through a succession of temporary jobs. At the same time, it is important to note that she was also privileged in several ways. She was an upper-middle-class White woman being educated at a time of public outcries for the educational rights of women. She did not fit the traditional role of a woman at a time that these roles were being challenged and champions were being sought. Senechal provides the pieces for a classic illustration of a woman's experiences in the culture of science; yet the discussion is not fully developed.

I learned much from this multifaceted, multidirectional view into the life of Dr. Dorothy Wrinch and the many individuals in her life. I learned in a nontraditional manner, which is probably fitting given the subject herself. As a science educator, I believe this book offers something that I have not readily found elsewhere—a female scientist's life story explored *within* the culture of society and science. Specifically, I believe that the book offers ideas and understandings that could lay the necessary foundation for a scholarly dialogue in a seminar course on gender, culture, and science. The multiple stories of scientists and mathematicians coming together in the context of time and cultures provides a type of resource I have not found elsewhere.

REFERENCES

- Delpit, L. (1988). The silenced dialogue: Power and pedagogy in educating other people's children. *Harvard Educational Review*, 58(3), 280–298.
- Ulrich, L. (1976). Vertuous women found: New England Ministerial Literature, 1668–1735. *American Quarterly*, 28(1), 20–40.

GAYLE A. BUCK

*Department of Curriculum &
Instruction*

Indiana University

Bloomington, IN 47405, USA

DOI 10.1002/sce.21103

Published online 30 January 2014 in Wiley Online Library (wileyonlinelibrary.com).

WILEY Job Network

WE MAKE YOUR RESEARCH EASY.
NOW WE MAKE JOB HUNTING EASY.



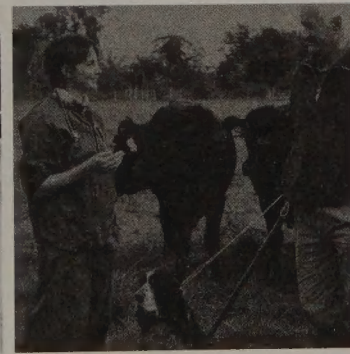
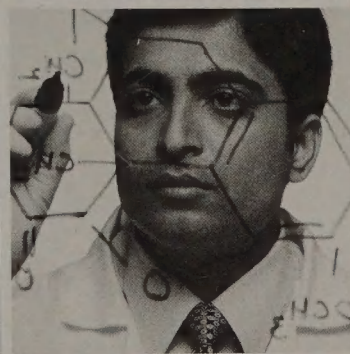
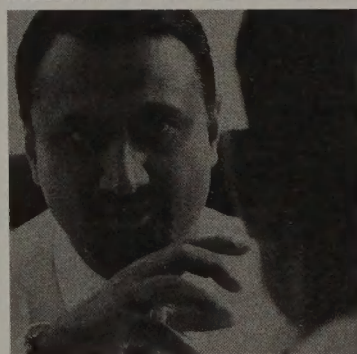
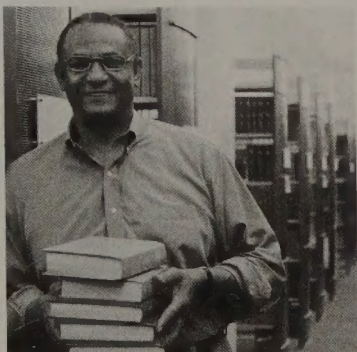
Let your partners in research energize your career.

Drawing on our expertise and relationships across the research and business communities, Wiley-Blackwell invites you to join Wiley Job Network, the definitive job site for professionals in the sciences, technology, business, finance, healthcare and the arts.

- **FIND** premium jobs from the most respected names in your industry
- **ATTRACT** hundreds of recruiters and employers in your field
- **CREATE** job alerts that match your criteria
- **OBTAIN** expert career advice and candidate resources

Register and upload your resume/CV now to begin your job search!

wileyjobnetwork.com



WILEY

exchanges.wiley.com/societies

Decision Sciences Journal of Innovative Education

"Their support is strong, continuous, enthusiastic, prompt, and supports the long-term growth of the journal and the institute."

Chetan S. Sankar,
Editor

Institute of Food Technologists

"By combining excellent customer service and a strategic view of the publishing market, they have enabled our scientific society to meet our goals..."

Jerry Bowman,
Vice President of
Communication

Veterinary Clinical Pathology

"I continue to be impressed with both the day-to-day management, including careful attention to quality issues, and long-term planning. We look forward to continuing our relationship for many years..."

Karen M. Young,
Editor-in-Chief

Anywhere Article.

Any format, any device, any time.

Today, more than ever, we need access to information that is immediate, clear and communicable. As a member of the community we serve, you will know how important it is to have access to that data, whenever you need it, and wherever you are.

What is Anywhere Article?

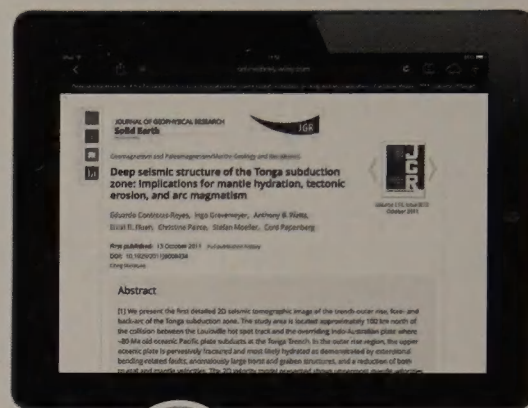
Anywhere Article is focused on making our online journal content on Wiley Online Library more readable and portable, whilst also allowing rich information to be brought to the surface. It achieves these goals in the following ways:

1. Readability

Clean design. Superfluous information and unnecessary distractions have been removed so that readers can focus on the article. Figures can be viewed in context or separately, and easily navigated, browsed or downloaded.

3. Mobility

Whatever device you use - desktop, tablet, or mobile - the article will be presented to take best advantage of that device, always readable, always easy to use, wherever you are.



2. Functionality

The new layout and sidebar tray allow readers access to important information, ie; references, figures, publication history at any point in the reading experience, without losing their place on the main page.

When Can I Start Using Anywhere Article?

☰ Enhanced Article (HTML) You can view an article in the new 'Anywhere Article' format wherever you see this link. You'll be able to view it easily on the device of your choice, at your convenience.

Visit www.wileyonlinelibrary.com today and look out for the new links underneath each journal article, try it, and see the difference for yourself.

WILEY

Wiley Online Library